MARK M. WILDE

FROM CLASSICAL TO QUANTUM SHANNON THEORY

MARK M. WILDE

FROM CLASSICAL TO QUANTUM SHANNON THEORY

Contents

I Introduction **15**

1 Concepts in Quantum Shannon Theory **17**
1.1 Overview of the Quantum Theory 20
1.2 The Emergence of Quantum Shannon Theory 24

2 Classical Shannon Theory **39**
2.1 Data Compression 39
2.2 Channel Capacity 48
2.3 Summary 62

II The Quantum Theory **65**

3 The Noiseless Quantum Theory **67**
3.1 Overview 68
3.2 Quantum Bits 68
3.3 Reversible Evolution 74
3.4 Measurement 81
3.5 Composite Quantum Systems 87
3.6 Summary and Extensions to Qudit States 101
3.7 History and Further Reading 108

4 The Noisy Quantum Theory **109**
4.1 Noisy Quantum States 110
4.2 Measurement in the Noisy Quantum Theory 122
4.3 Composite Noisy Quantum Systems 123
4.4 Noisy Evolution 132
4.5 Summary 151
4.6 History and Further Reading 152

5 The Purified Quantum Theory **153**
5.1 Purification . . . 154
5.2 Isometric Evolution . . . 155
5.3 Coherent Quantum Instrument . . . 165
5.4 Coherent Measurement . . . 166
5.5 History and Further Reading . . . 167

III Unit Quantum Protocols **169**

6 Three Unit Quantum Protocols **171**
6.1 Nonlocal Unit Resources . . . 172
6.2 Protocols . . . 174
6.3 Optimality of the Three Unit Protocols . . . 183
6.4 Extensions for Quantum Shannon Theory . . . 185
6.5 Three Unit Qudit Protocols . . . 186
6.6 History and Further Reading . . . 192

7 Coherent Protocols **193**
7.1 Definition of Coherent Communication . . . 194
7.2 Implementations of a Coherent Bit Channel . . . 197
7.3 Coherent Dense Coding . . . 198
7.4 Coherent Teleportation . . . 200
7.5 The Coherent Communication Identity . . . 202
7.6 History and Further Reading . . . 203

8 The Unit Resource Capacity Region **205**
8.1 The Unit Resource Achievable Region . . . 205
8.2 The Direct Coding Theorem . . . 210
8.3 The Converse Theorem . . . 210
8.4 History and Further Reading . . . 214

IV Tools of Quantum Shannon Theory **215**

9 Distance Measures **217**
9.1 Trace Distance . . . 218
9.2 Fidelity . . . 226
9.3 Relationships between Trace Distance and Fidelity . . . 233
9.4 Gentle Measurement . . . 237
9.5 Fidelity of a Noisy Quantum Channel . . . 240
9.6 The Hilbert-Schmidt Distance Measure . . . 244
9.7 History and Further Reading . . . 244

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

10 Classical Information and Entropy **247**
10.1 Entropy of a Random Variable . . . 248
10.2 Conditional Entropy . . . 252
10.3 Joint Entropy . . . 253
10.4 Mutual Information . . . 254
10.5 Relative Entropy . . . 255
10.6 Conditional Mutual Information . . . 256
10.7 Information Inequalities . . . 257
10.8 Classical Information and Entropy of Quantum Systems . . . 263
10.9 History and Further Reading . . . 266

11 Quantum Information and Entropy **267**
11.1 Quantum Entropy . . . 268
11.2 Joint Quantum Entropy . . . 272
11.3 Potential yet Unsatisfactory Definitions of Conditional Quantum Entropy . . . 276
11.4 Conditional Quantum Entropy . . . 278
11.5 Coherent Information . . . 280
11.6 Quantum Mutual Information . . . 282
11.7 Conditional Quantum Mutual Information . . . 285
11.8 Quantum Relative Entropy . . . 286
11.9 Quantum Information Inequalities . . . 289
11.10History and Further Reading . . . 304

12 The Information of Quantum Channels **307**
12.1 Mutual Information of a Classical Channel . . . 308
12.2 Private Information of a Wiretap Channel . . . 314
12.3 Holevo Information of a Quantum Channel . . . 317
12.4 Mutual Information of a Quantum Channel . . . 324
12.5 Coherent Information of a Quantum Channel . . . 329
12.6 Private Information of a Quantum Channel . . . 334
12.7 Summary . . . 340
12.8 History and Further Reading . . . 340

13 Classical Typicality **343**
13.1 An Example of Typicality . . . 344
13.2 Weak Typicality . . . 345
13.3 Properties of the Typical Set . . . 347
13.4 Application of Typical Sequences: Shannon Compression . . . 349
13.5 Weak Joint Typicality . . . 351
13.6 Weak Conditional Typicality . . . 354
13.7 Strong Typicality . . . 356
13.8 Strong Joint Typicality . . . 366
13.9 Strong Conditional Typicality . . . 367

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

13.10Application: Shannon's Channel Capacity Theorem 373
13.11Concluding Remarks 378
13.12History and Further Reading 378

14 Quantum Typicality **379**
14.1 The Typical Subspace 380
14.2 Conditional Quantum Typicality 389
14.3 The Method of Types for Quantum Systems 399
14.4 Concluding Remarks 402
14.5 History and Further Reading 402

15 The Packing Lemma **403**
15.1 Introductory Example 404
15.2 The Setting of the Packing Lemma 404
15.3 Statement of the Packing Lemma 406
15.4 Proof of the Packing Lemma 408
15.5 Derandomization and Expurgation 413
15.6 History and Further Reading 415

16 The Covering Lemma **417**
16.1 Introductory Example 418
16.2 Setting and Statement of the Covering Lemma 420
16.3 Proof of the Covering Lemma 422
16.4 History and Further Reading 429

V Noiseless Quantum Shannon Theory **431**

17 Schumacher Compression **433**
17.1 The Information Processing Task 434
17.2 The Quantum Data Compression Theorem 435
17.3 Quantum Compression Example 440
17.4 Variations on the Schumacher Theme 441
17.5 Concluding Remarks 442
17.6 History and Further Reading 443

18 Entanglement Concentration **445**
18.1 An Example of Entanglement Concentration 446
18.2 The Information Processing Task 449
18.3 The Entanglement Concentration Theorem 449
18.4 Common Randomness Concentration 455
18.5 Schumacher Compression versus Entanglement Concentration 457
18.6 Concluding Remarks 460
18.7 History and Further Reading 460

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

VI Noisy Quantum Shannon Theory 463

19 Classical Communication 467
19.1 Naive Approach: Product Measurements at the Decoder 469
19.2 The Information Processing Task 471
19.3 The Classical Capacity Theorem 474
19.4 Examples of Channels 479
19.5 Superadditivity of the Holevo Information 487
19.6 Concluding Remarks 490
19.7 History and Further Reading 491

20 Entanglement-Assisted Classical Communication 493
20.1 The Information Processing Task 495
20.2 A Preliminary Example 496
20.3 The Entanglement-Assisted Classical Capacity Theorem 500
20.4 The Direct Coding Theorem 500
20.5 The Converse Theorem 509
20.6 Examples of Channels 517
20.7 Concluding Remarks 522
20.8 History and Further Reading 523

21 Coherent Communication with Noisy Resources 525
21.1 Entanglement-Assisted Quantum Communication 526
21.2 Quantum Communication 531
21.3 Noisy Super-Dense Coding 532
21.4 State Transfer 535
21.5 Trade-off Coding 538
21.6 Concluding Remarks 546
21.7 History and Further Reading 547

22 Private Classical Communication 549
22.1 The Information Processing Task 550
22.2 The Private Classical Capacity Theorem 552
22.3 The Direct Coding Theorem 553
22.4 The Converse Theorem 562
22.5 Discussion of Private Classical Capacity 563
22.6 History and Further Reading 565

23 Quantum Communication 567
23.1 The Information Processing Task 568
23.2 The No-Cloning Theorem and Quantum Communication 570
23.3 The Quantum Capacity Theorem 571
23.4 The Direct Coding Theorem 571

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

23.5 Converse Theorem . . . 579
23.6 Example Channels . . . 581
23.7 Discussion of Quantum Capacity . . . 584
23.8 Entanglement Distillation . . . 589
23.9 History and Further Reading . . . 592

24 Trading Resources for Communication 595
24.1 The Information Processing Task . . . 596
24.2 The Quantum Dynamic Capacity Theorem . . . 598
24.3 The Direct Coding Theorem . . . 603
24.4 The Converse Theorem . . . 605
24.5 Examples of Channels . . . 616
24.6 History and Further Reading . . . 624

25 Summary and Outlook 627
25.1 Unit Protocols . . . 628
25.2 Noiseless Quantum Shannon Theory . . . 629
25.3 Noisy Quantum Shannon Theory . . . 630
25.4 Protocols not covered in this book . . . 632
25.5 Network Quantum Shannon Theory . . . 633
25.6 Future Directions . . . 634

A Miscellaneous Mathematics 635
A.1 The Operator Chernoff Bound . . . 638

B Monotonicity of Quantum Relative Entropy 643

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

How to use this book

For students

Prerequisites for understanding the content in this book are a solid background in probability theory and linear algebra. If you are new to information theory, then there is enough background in this book to get you up to speed (Chapters 2, 10, 12, and 13). Though, classics on information theory such as Cover and Thomas [57] and MacKay [189] could be helpful as a reference. If you are new to quantum mechanics, then there should be enough material in this book (Part II) to give you the background necessary for understanding quantum Shannon theory. The book of Nielsen and Chuang (sometimes known as "Mike and Ike") has become the standard starting point for students in quantum information science and might be helpful as well [197]. Some of the content in that book is available in Nielsen's dissertation [194]. If you are familiar with Shannon's information theory (at the level of Cover and Thomas [57], for example), then this book should be a helpful entry point into the field of quantum Shannon theory. We build on intuition developed classically to help in establishing schemes for communication over quantum channels. If you are familiar with quantum mechanics, it might still be worthwhile to review Part II because some content there might not be part of a standard course on quantum mechanics.

The aim of this book is to develop "from the ground up" many of the major, exciting, pre- and post-millenium developments in the general area of study known as quantum Shannon theory. As such, we spend a significant amount of time on quantum mechanics for quantum information theory (Part II), we give a careful study of the important unit protocols of teleportation, super-dense coding, and entanglement distribution (Part III), and we develop many of the tools necessary for understanding information transmission or compression (Part IV). Parts V and VI are the culmination of this book, where all of the tools developed come into play for understanding many of the important results in quantum Shannon theory.

For instructors

This book could be useful for self-learning or as a reference, but one of the main goals is for it to be employed as an instructional aid for the classroom. To aid instructors in designing a course to suit their own needs, this book is available under a Creative Commons Attribution-NonCommercial-ShareAlike license. This means that you can modify and redistribute the

book as you wish, as long as you attribute the author of this book, you do not use it for commercial purposes, and you share a modification or derivative work under the same license (see `http://creativecommons.org/licenses/by-nc-sa/3.0/` for a readable summary of the terms of the license). These requirements can be waived if you obtain permission from the present author. By releasing the book under this license, I expect and encourage instructors to modify this book for their own needs. This will allow for the addition of new exercises, new developments in the theory, and the latest open problems. It might also be a helpful starting point for a book on a related topic, such as network quantum Shannon theory.

I used an earlier version of this book in a one-semester course on quantum Shannon theory at McGill University during Winter semester 2011 (in many parts of the US, this semester is typically called "Spring semester"). We almost went through the entire book, but it might also be possible to spread the content over two semesters instead. Here is the order in which we proceeded:

1. Introduction in Part I
2. Quantum mechanics in Part II
3. Unit protocols in Part III
4. Chapter 9 on distance measures, Chapter 10 on classical information and entropy, and Chapter 11 on quantum information and entropy.
5. The first part of Chapter 13 on classical typicality and Shannon compression.
6. The first part of Chapter 14 on quantum typicality.
7. Chapter 17 on Schumacher compression.
8. Back to Chapters 13 and 14 for the method of types.
9. Chapter 18 on entanglement concentration.
10. Chapter 19 on classical communication.
11. Chapter 20 on entanglement-assisted classical communication.
12. The final explosion of results in Chapter 21 (one of which is a route to proving the achievability part of the quantum capacity theorem).

The above order is just a particular order that suited the needs for the class at McGill, but other orders are of course possible. One could sacrifice the last part of Part III on the unit resource capacity region if there is no desire to cover the quantum dynamic capacity theorem. One could also focus on going from classical communication to private classical communication to quantum communication in order to develop some more intuition behind the quantum capacity theorem.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Other sources

There are many other sources to obtain a background in quantum Shannon theory. The standard reference has become the book of Nielsen and Chuang [197], but it does not feature any of the post-millenium results in quantum Shannon theory. Other books that cover some aspects of quantum Shannon theory are Hayashi's [128] and Holevo's [145]. Patrick Hayden has had a significant hand as a collaborative guide for many PhD and Masters' theses in quantum Shannon theory, during his time as a postdoctoral fellow at the California Institute of Technology and as a professor at McGill University. These include the theses of Yard [266], Abeyesinghe [2], Savov [210, 211], Dupuis [82], and Dutil [85]. All of these theses are excellent references. Naturally, Hayden also had a strong influence over the present author during the development of this book.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Acknowledgments

I began working on this book in the summer of 2008 in Los Angeles, with much time to spare in the final months of dissertation writing. I had a strong determination to review quantum Shannon theory, a beautiful area of quantum information science that Igor Devetak had taught me three years earlier at USC in fall 2005. I was carefully studying a manuscript entitled "Principles of Quantum Information Theory," a text that Igor had initiated in collaboration with Patrick Hayden and Andreas Winter. I read this manuscript many times, and many parts of it I understood well, though other parts I did not.

After a few weeks of reading and rereading, I decided "if I can write it out myself from scratch, perhaps I would then understand it!", and thus began the writing of the chapters on the Packing Lemma, the Covering Lemma, and quantum typicality. I knew that Igor's (now former) students Min-Hsiu Hsieh and Zhicheng Luo knew the topic well because they had already written several quality research papers with him, so I requested if they could meet with me weekly for an hour to review the fundamentals. They kindly agreed and helped me quite a bit in understanding the packing and covering techniques.

Not much later, after graduating, I began collaborating with Min-Hsiu on a research project that Igor had suggested to the both of us: "find the triple trade-off capacity formulas of a quantum channel." This was perhaps the best starting point for me to learn quantum Shannon theory because proving this theorem required an understanding of most everything that had already been accomplished in the area. After a month of effort, I continued to work with Min-Hsiu on this project while joining Andreas Winter's Singapore group for a two-month visit. As I learned more, I added more to the notes, and they continued to grow.

After landing a job in the DC area for January 2009, I realized that I had almost enough material for teaching a course, and so I contacted local universities in the area to see if they would be interested. Can Korman, formerly chair of the Electrical Engineering Department at George Washington University, was excited about the possibility. His enthusiasm was enough to keep me going on the notes, and so I continued to refine and add to them in my spare time in preparing for teaching. Unfortunately (or perhaps fortunately?), the course ended up being canceled. This was disheartening to me, but in the mean time, I had contacted Patrick Hayden to see if he would be interested in having me join his group at McGill University for postdoctoral studies. Patrick Hayden and David Avis then offered me a postdoctoral fellowship, and I moved to Montreal in October 2009.

After joining, I learned a lot by collaborating and discussing with Patrick and his group members. Patrick offered me the opportunity to teach his graduate class on quantum Shan-

non theory while he was away on sabbatical, and this encouraged me further to persist with the notes.

I am grateful to everyone mentioned above for encouraging and supporting me during this project, and I am also grateful to everyone who provided feedback during the course of writing up. In this regard, I am especially grateful to Dave Touchette for detailed feedback on all of the chapters in the book. Dave's careful reading and spotting of errors has immensely improved the quality of the book. I am grateful to my father, Gregory E. Wilde, Sr., for feedback on earlier chapters and for advice and love throughout. I thank Ivan Savov for encouraging me, for feedback, and for believing that this is an important scholarly work. I also thank Constance Caramanolis, Raza-Ali Kazmi, John M. Schanck, Bilal Shaw, and Anna Vershynina for valuable feedback. I am grateful to Min-Hsiu Hsieh for the many research topics we have worked on together that have enhanced my knowledge of quantum Shannon theory. I thank Michael Nielsen and Victor Shoup for advice on Creative Commons licensing and Kurt Jacobs for advice on book publishing. I acknowledge funding from the MDEIE (Quebec) PSR-SIIRI international collaboration grant.

I am indebted to my mentors who took me on as a student during my career. Todd Brun was a wonderful PhD supervisor—helpful, friendly, and encouraging of creativity and original pursuit. Igor Devetak taught me quantum Shannon theory in fall 2005 and helped me once per week during his office hours. He also invited me to join Todd's and his group, and more recently, Igor provided much encouragement and "big-picture" feedback during the writing of this book. Bart Kosko shaped me as a scholar during my early years at USC and provided helpful advice regarding the book project. Patrick Hayden has been an immense bedrock of support at McGill. His knowledge of quantum information and many other areas is unsurpassed, and he has been kind, inviting, and helpful during my time at McGill. I am also grateful to Patrick for giving me the opportunity to teach at McGill and for advice throughout the development of this book.

I thank my mother, father, sister, and brother and all of my surrounding family members for being a source of love and support. Finally, I am indebted to Christabelle and her family for warmth and love. I dedicate this book to the memory of my grandparents Joseph and Rose McMahon, and Norbert Jay and Mary Wilde. *Lux aeterna luceat eis, Domine.*

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Part I

Introduction

CHAPTER 1

Concepts in Quantum Shannon Theory

In these first few chapters, our aim is to establish a firm grounding so that we can address some fundamental questions regarding information transmission over quantum channels. This area of study has become known as "quantum Shannon theory" in the broader quantum information community, in order to distinguish this topic from other areas of study in quantum information science. In this text, we will use the terms "quantum Shannon theory" and "quantum information theory" somewhat interchangeably. We will begin by briefly overviewing several fundamental aspects of the quantum theory. Our study of the quantum theory, in this chapter and future ones, will be at an abstract level, without giving preference to any particular physical system such as a spin-1/2 particle or a photon. This approach will be more beneficial for the purposes of our study, but, here and there, we will make some reference to actual physical systems to ground us in reality.

You may be wondering, what is *quantum Shannon theory* and why do we name this area of study as such? In short, quantum Shannon theory is the study of the ultimate capability of noisy physical systems, governed by the laws of quantum mechanics, to preserve information and correlations. Quantum information theorists have chosen the name *quantum Shannon theory* to honor Claude Shannon, who single-handedly founded the field of classical information theory, with a groundbreaking 1948 paper [222]. In particular, the name refers to the asymptotic theory of quantum information, which is the main topic of study in this book. Information theorists since Shannon have dubbed him the "Einstein of the information age."[1] The name *quantum Shannon theory* is fit to capture this area of study because we use quantum versions of Shannon's ideas to prove some of the main theorems in quantum Shannon theory.

We prefer the title "quantum Shannon theory" over such titles as "quantum information science" or just "quantum information." These other titles are too broad, encompassing subjects as diverse as quantum computation, quantum algorithms, quantum complexity the-

[1] It is worthwhile to look up "Claude Shannon—Father of the Information Age" on YouTube and watch several reknowned information theorists speak with awe about "the founding father" of information theory.

ory, quantum communication complexity, entanglement theory, quantum key distribution, quantum error correction, and even the experimental implementation of quantum protocols. Quantum Shannon theory does overlap with some of the aforementioned subjects, such as quantum computation, entanglement theory, quantum key distribution, and quantum error correction, but the name "quantum Shannon theory" should evoke a certain paradigm for quantum communication with which the reader will become intimately familiar after some exposure to the topics in this book. For example, it is necessary for us to discuss *quantum gates* (a topic in quantum computing) because quantum Shannon-theoretic protocols exploit them to achieve certain information processing tasks. Also, in Chapter 22, we are interested in the ultimate limitation on the ability of a noisy quantum communication channel to transmit private information (information that is secret from any third party besides the intended receiver). This topic connects quantum Shannon theory with quantum key distribution because the private information capacity of a noisy quantum channel is strongly related to the task of using the quantum channel to distribute a secret key. As a final connection, perhaps the most important theorem of quantum Shannon theory is the *quantum capacity theorem*. This theorem determines the ultimate rate at which a quantum channel can reliably transmit quantum information to a receiver. The result provided by the quantum capacity theorem is closely related to the theory of quantum error correction, but the mathematical techniques used in quantum Shannon theory and in quantum error correction are so different that these subjects merit different courses of study.

Quantum Shannon theory intersects two of the great sciences of the twentieth century: the quantum theory and information theory. It was really only a matter of time before physicists, mathematicians, computer scientists, and engineers began to consider the convergence of the two subjects because the quantum theory was essentially established by 1926 and information theory by 1948. This convergence has sparked what we may call the "quantum information revolution" or what some refer to as the "second quantum revolution" [81] (with the first one being the discovery of the quantum theory).

The fundamental components of the quantum theory are a set of postulates that govern phenomena on the scale of atoms. Uncertainty is at the heart of the quantum theory—"quantum uncertainty" or "Heisenberg uncertainty" is not due to our lack or loss of information or due to imprecise measurement capability, but rather, it is a fundamental uncertainty inherent in nature itself. The discovery of the quantum theory came about as a total shock to the physics community, shaking the foundations of scientific knowledge. Perhaps it is for this reason that every introductory quantum mechanics course delves into its history in detail and celebrates the founding fathers of the quantum theory. In this book, we do not discuss the history of the quantum theory in much detail and instead refer to several great introductory books for these details [95, 41, 209, 116]. Physicists such as Planck, Einstein, Bohr, de Broglie, Born, Heisenberg, Schrödinger, Pauli, Dirac, and von Neumann contributed to the foundations of the quantum theory in the 1920s and 1930s. We introduce the quantum theory by *briefly* commenting on its history and major underlying concepts.

Information theory is the second great foundational science for quantum Shannon theory. In some sense, it is merely an application of probability theory. Its aim is to quantify the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

ultimate compressibility of information and the ultimate ability for a sender to transmit information reliably to a receiver. It relies upon probability theory because a "classical" uncertainty, arising from our lack of total information about any given scenario, is ubiquitous throughout all information processing tasks. The uncertainty in classical information theory is the kind that is present in the flipping of a coin or the shuffle of a deck of cards, the uncertainty due to imprecise knowledge. "Quantum" uncertainty is inherent in nature itself and is perhaps not as intuitive as the uncertainty that classical information theory measures. We later expand further on these differing kinds of uncertainty, and Chapter 4 shows how a theory of quantum information captures both kinds of uncertainty within one formalism.[2]

The history of classical information theory began with Claude Shannon. Shannon's contribution is heralded as one of the single greatest contributions to modern science because he established the entire field in his seminal 1948 paper [222]. In this paper, he coined the essential terminology, and he stated and justified the main mathematical definitions and the two fundamental theorems of information theory. Many successors have contributed to information theory, but most, if not all, of the follow-up contributions employ Shannon's line of thinking in some form. In quantum Shannon theory, we will notice that many of Shannon's original ideas are present, though they take a particular "quantum" form.

One of the major assumptions in both classical information theory and quantum Shannon theory is that local computation is free but communication is expensive. In particular, for the classical case, we assume that each party has unbounded computation available. For the quantum case, we assume that each party has a fault-tolerant quantum computer available at his or her local station and the power of each quantum computer is unbounded. We also assume that both communication and a shared resource are expensive, and for this reason, we keep track of these resources in a *resource count*. Though sometimes, we might say that classical communication is free in order to simplify a scenario. A simplification like this one can lead to greater insights that would not be possible without making such an assumption.

We should first study and understand the postulates of the quantum theory in order to study quantum Shannon theory properly. Your heart may sink when you learn that the Nobel-prize winning physicist Richard Feynman is famously quoted as saying, "I think I can safely say that nobody understands quantum mechanics." We should clarify Feynman's statement. Of course, Feynman does not intend to suggest that no one knows how to work with the quantum theory. Many well-abled physicists are employed to spend their days exploiting the laws of the quantum theory to do fantastic things, such as the trapping of ions in a vacuum or applying the quantum tunneling effect in a transistor to process a single electron. I am hoping that you will give me the license to interpret Feynman's statement. I think he means that it is impossible for us to understand the quantum theory intuitively because we do not experience the phenomena that it predicts. If we were the size of atoms and we experienced the laws of quantum theory on a daily basis, then perhaps the quantum theory would be as intuitive to us as Newton's law of universal gravitation.[3] Thus, in this

[2]Von Neumann established the density matrix formalism in his 1932 book on the quantum theory. This mathematical framework captures both kinds of uncertainty [243].

[3]Of course, Newton's law of universal gravitation was a revolutionary breakthrough because the phe-

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

sense, I would agree with Feynman—nobody can really understand the quantum theory because it is not part of our every day experiences. Nevertheless, our aim in this book is to work with the laws of quantum theory so that we may begin to gather insights about what the theory predicts. Only by exposure to and practice with its postulates can we really gain an intuition for its predictions. It is best to imagine that the world in our every day life does incorporate the postulates of quantum mechanics, because, indeed, as many, many experiments have confirmed, it does!

We delve into the history of the convergence of the quantum theory and information theory in some detail in this introductory chapter because this convergence does have an interesting history and is relevant to the topic of this book. The purpose of this historical review is not only to become familiar with the field itself but also to glimpse into the minds of the founders of the field so that we may see the types of questions that are important to think about when tackling new, unsolved problems.[4] Many of the most important results come about from asking simple, yet profound, questions and exploring the possibilities.

We first briefly review the history and the fundamental concepts of the quantum theory before delving into the convergence of the quantum theory and information theory. We build on these discussions by introducing some of the initial fundamental contributions to quantum Shannon theory. The final part of this chapter ends by posing some of the questions to which quantum Shannon theory provides answers.

1.1 Overview of the Quantum Theory

1.1.1 Brief History of the Quantum Theory

A physicist living around 1890 would have been well pleased with the progress of physics, but perhaps frustrated at the seeming lack of open research problems. It seemed as though the Newtonian laws of mechanics, Maxwell's theory of electromagnetism, and Boltzmann's theory of statistical mechanics explained most natural phenomena. In fact, Max Planck, one of the founding fathers of the quantum theory, was searching for an area of study in 1874 and his advisor gave him the following guidance:

> "In this field [of physics], almost everything is already discovered, and all that remains is to fill a few holes."

nomenon of gravity is not entirely intuitive when a student first learns it. But, we do experience the gravitational law in our daily lives and I would argue that this phenomenon is much more intuitive than, say, the phenomenon of quantum entanglement.

[4] Another way to discover good questions is to attend parties that well-established professors hold. The story goes that Oxford physicist David Deutsch attended a 1981 party at the Austin, Texas house of renowned physicist John Archibald Wheeler, in which many attendees discussed the foundations of computing [193]. Deutsch claims that he could immediately see that the quantum theory would give an improvement for computation. Not quite immediately later, he published an algorithm in 1985 that was the first instance of a quantum speedup over the fastest classical algorithm [67].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Two Clouds

Fortunately for the quantum theory, Planck did not heed this advice and instead began his physics studies. Not everyone agreed with Planck's former advisor. Lord Kelvin stated in his famous April 1900 lecture that "two clouds" surrounded the "beauty and clearness of theory" [1]. The first cloud was the failure of Michelson and Morley to detect a change in the speed of light as predicted by an "ether theory," and the second cloud was the ultraviolet catastrophe, the prediction of classical theory that a blackbody emits radiation with an infinite intensity at high ultraviolet frequencies. In that same year of 1900, Planck started the quantum revolution that began to clear the second cloud. He assumed that light comes in discrete bundles of energy and used this idea to produce a formula that correctly predicts the spectrum of blackbody radiation [205]. A great cartoon lampoon of the ultraviolet catastrophe shows Planck calmly sitting fireside with a classical physicist whose face is burning to bits because of the intense ultraviolet radiation that his classical theory predicts the fire is emitting [190]. A few years later, in 1905, Einstein contributed a paper that helped to further clear the second cloud [86] (he also cleared the first cloud with his other 1905 paper on special relativity). He assumed that Planck was right and showed that the postulate that light arrives in "quanta" (now known as the photon theory) provides a simple explanation for the photoelectric effect, the phenomenon in which electromagnetic radiation beyond a certain threshold frequency impinging on a metallic surface induces a current in that metal.

These two explanations of Planck and Einstein fueled a theoretical revolution in physics that some now call the first quantum revolution [81]. Some years later, in 1924, Louis de Broglie postulated that every individual element of matter, whether an atom, electron, or photon, has both particle-like behavior and wave-like behavior [66]. Just two years later, Erwin Schrödinger used the de Broglie idea to formulate a wave equation, now known as Schrödinger's equation, that governs the evolution of a closed quantum-mechanical system [214]. His formalism later became known as wave mechanics and was popular among physicists because it appealed to notions with which they were already familiar. Meanwhile, in 1925, Werner Heisenberg formulated an "alternate" quantum theory called matrix mechanics [137]. His theory used matrices and theorems from linear algebra, mathematics with which physicists at the time were not readily familiar. For this reason, Schrödinger's wave mechanics was more popular than Heisenberg's matrix mechanics. In 1930, Paul Dirac published a textbook (now in its fourth edition and reprinted 16 times) that unified the formalisms of Schrödinger and Heisenberg, showing that they were actually equivalent [79]. In a later edition, he introduced the now ubiquitous "Dirac notation" for quantum theory that we will employ in this book.

After the publication of Dirac's textbook, the quantum theory then stood on firm mathematical grounding and the basic theory had been established. We thus end our historical overview at this point and move on to the fundamental concepts of the quantum theory.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

1.1.2 Fundamental Concepts of the Quantum Theory

Quantum theory, as applied in quantum information theory, really has only a few important concepts. We review each of these aspects of quantum theory briefly in this section. Some of these phenomena are uniquely "quantum" but others do occur in the classical theory. In short, these concepts are as follows[5]:

1. Indeterminism
2. Interference
3. Uncertainty
4. Superposition
5. Entanglement

The quantum theory is *indeterministic* because the theory makes predictions about probabilities of events only. This aspect of quantum theory is in contrast with a deterministic classical theory such as that predicted by the Newtonian laws. In the Newtonian system, it is possible to predict, with certainty, the trajectories of all objects involved in an interaction if one knows only the initial positions and velocities of all the objects. This deterministic view of reality even led some to believe in determinism from a philosophical point of view. For instance, the mathematician Pierre-Simon Laplace once stated that a supreme intellect, colloquially known as Laplace's demon, could predict all future events from present and past events:

> "We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."

The application of Laplace's statement to atoms is fundamentally incorrect, but we can forgive him because the quantum theory had not yet been established in his time. Many have extrapolated from Laplace's statement to argue the invalidity of human free will. We leave such debates to philosophers.[6]

In reality, we never can possess full information about the positions and velocities of every object in any given physical system. Incorporating probability theory then allows us to make predictions about the probabilities of events and, with some modifications, the

[5] I have used Todd A. Brun's list from his lecture notes [49].

[6] John Archibald Wheeler may disagree with this approach. He once said, "Philosophy is too important to be left to the philosophers" [191].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

classical theory becomes an indeterministic theory. Thus, indeterminism is not a unique aspect of the quantum theory but merely a feature of it. But this feature is so crucial to the quantum theory that we list it among the fundamental concepts.

Interference is another feature of the quantum theory. It is also present in any classical wave theory—constructive interference occurs when the crest of one wave meets the crest of another, producing a stronger wave, while destructive interference occurs when the crest of one wave meets the trough of another, canceling out each other. In any classical wave theory, a wave occurs as a result of many particles in a particular medium coherently displacing one another, as in an ocean surface wave or a sound pressure wave, or as a result of coherent oscillating electric and magnetic fields, as in an electromagnetic wave. The strange aspect of interference in the quantum theory is that even a single "particle" such as an electron can exhibit wave-like features, as in the famous double slit experiment (see Ref. [115] for a history of these experiments). This quantum interference is what contributes wave-particle duality to every fundamental component of matter.

Uncertainty is at the heart of the quantum theory. Uncertainty in the quantum theory is fundamentally different from uncertainty in the classical theory (discussed in the former paragraph about an indeterministic classical theory). The archetypal example of uncertainty in the quantum theory occurs for a single particle. This particle has two complementary variables: its position and its momentum. The uncertainty principle states that it is impossible to know both its position and momentum precisely. This principle even calls into question the meaning of the word "know" in the previous sentence in the context of quantum theory. We might say that we can only know that which we measure, and thus, we can only know the position of a particle after performing a precise measurement that determines it. If we follow with a precise measurement of its momentum, we lose all information about the position of the particle after learning its momentum. In quantum information science, the BB84 protocol for quantum key distribution exploits the uncertainty principle and statistical analysis to determine the presence of an eavesdropper on a quantum communication channel by encoding information into two complementary variables [22].

The *superposition* principle states that a quantum particle can be in a linear combination state, or *superposed state*, of any two other allowable states. This principle is a result of the linearity of quantum theory. Schrodinger's wave equation is a linear differential equation, meaning that the linear combination $\alpha\psi+\beta\phi$ is a solution of the equation if ψ and ϕ are both solutions of the equation. We say that the solution $\alpha\psi+\beta\phi$ is a coherent superposition of the two solutions. The superposition principle has dramatic consequences for the interpretation of the quantum theory—it gives rise to the notion that a particle can somehow "be in one location and another" at the same time. There are different interpretations of the meaning of the superposition principle, but we do not highlight them here. We merely choose to use the technical language that the particle is in a superposition of both locations. The loss of a superposition can occur through the interaction of a particle with its environment. Maintaining an arbitrary superposition of quantum states is one of the central goals of a quantum communication protocol.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The last, and perhaps most striking, "quantum" feature that we highlight here is *entanglement*. There is no true classical analog of entanglement. The closest analog of entanglement might be a secret key that two parties possess, but even this analogy does not come close. Entanglement refers to the strong quantum correlations that two or more quantum particles can possess. The correlations in quantum entanglement are stronger than any classical correlations in a precise, technical sense. Schrödinger first coined the term "entanglement" after observing some of its strange properties and consequences [215]. Einstein, Podolsky, and Rosen then presented an apparent paradox involving entanglement that raised concerns over the completeness of the quantum theory [87]. That is, they suggested that the seemingly strange properties of entanglement called the uncertainty principle into question (and thus the completeness of the quantum theory) and furthermore suggested that there might be some "local hidden-variable" theory that could explain the results of experiments. It took about thirty years to resolve this paradox, but John Bell did so by presenting a simple inequality, now known as a Bell inequality [17]. He showed that any two-particle classical correlations that satisfy the assumptions of the "local hidden-variable theory" of Einstein, Podolsky, and Rosen must be less than a certain amount. He then showed how the correlations of two entangled quantum particles can violate this inequality, and thus, entanglement has no explanation in terms of classical correlations but is instead a uniquely quantum phenomenon. Experimentalists later verified that two entangled quantum particles can violate Bell's inequality [10].

In quantum information science, the non-classical correlations in entanglement play a fundamental role in many protocols. For example, entanglement is the enabling resource in teleportation, a protocol that disembodies a quantum state in one location and reproduces it in another. We will see many other examples of exploiting entanglement throughout this book.

Entanglement theory concerns different methods for quantifying the amount of entanglement present not only in a two-particle state but also in a multiparticle state. A large body of literature exists that investigates entanglement theory [155], but we only address aspects of entanglement that are relevant in our study of quantum Shannon theory.

The above five features capture the essence of the quantum theory, but we will see more aspects of it as we progress through our overview in Chapters 3, 4, and 5.

1.2 The Emergence of Quantum Shannon Theory

In the previous section, we discussed several unique quantum phenomena such as superposition and entanglement, but it is not clear what kind of information these unique quantum phenomena represent. Is it possible to find a convergence of the quantum theory and Shannon's information theory, and if so, what is the convergence?

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

1.2.1 The Shannon Information Bit

A fundamental contribution of Shannon is the notion of a *bit* as a measure of information. Typically, when we think of a bit, we think of a two-valued quantity that can be in the state 'off' or the state 'on.' We represent this bit with a binary number that can be '0' or '1.' We also associate a physical representation with a bit—this physical representation can be whether a light switch is off or on, whether a transistor allows current to flow or does not, whether a large number of magnetic spins point in one direction or another, the list going on and on. These are all physical notions of a bit.

Shannon's notion of a bit is quite different from these physical notions, and we motivate his notion with the example of a fair coin. Without flipping the coin, we have no idea what the result of a coin flip will be—our best guess at the result is to guess randomly. If someone else learns the result of a random coin flip, we can ask this person the question: What was the result? We then learn *one bit of information*.

Though it may seem obvious, it is important to stress that we do not learn any or not as much information if we do not ask the right question. This point becomes even more important in the quantum case. Suppose that the coin is not fair—without loss of generality, suppose the probability of "heads" is greater than the probability of "tails." In this case, we would not be as surprised to learn that the result of a coin flip is "heads." We may say in this case that we learn less than one bit of information if we were to ask someone the result of the coin flip.

The Shannon binary entropy is a measure of information. Given a probability distribution $(p, 1-p)$ for a binary random variable, its Shannon binary entropy is

$$H_2(p) \equiv -p \log p - (1-p) \log(1-p), \tag{1.1}$$

where the logarithm is base two. The Shannon binary entropy measures information in units of bits. We will discuss it in more detail in the next chapter and in Chapter 10.

The Shannon bit, or Shannon binary entropy, is a measure of the surprise upon learning the outcome of a random binary experiment. Thus, the Shannon bit has a completely different interpretation from that of the physical bit. The outcome of the coin flip resides in a physical bit, but it is the information associated with the random nature of the physical bit that we would like to measure. It is this notion of bit that is important in information theory.

1.2.2 A Measure of Quantum Information

The above section discusses Shannon's notion of a bit as a measure of information. A natural question is whether there is an analogous measure of quantum information, but before we can even ask that question, we might first wonder: what is *quantum information*? As in the classical case, there is a *physical* notion of quantum information. A quantum state always resides "in" a physical system. Perhaps another way of stating this idea is that every physical system is in some quantum state. The physical notion of a quantum bit, or qubit for short (pronounced "cue $\cdot$ bit"), is a two-level quantum system. Examples of two-level quantum

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

systems are the spin of the electron, the polarization of a photon, or an atom with a ground state and an excited state. The physical notion of a qubit is straightforward to understand once we have a grasp of the quantum theory.

A more pressing question for us in this book is to understand an *informational* notion of a qubit, as in the Shannon sense. In the classical case, we quantify information by the amount of knowledge we gain after learning the answer to a probabilistic question. In the quantum world, what knowledge can we have of a quantum state?

Sometimes we may know the exact quantum state of a physical system because we prepared the quantum system in a certain way. For example, we may prepare an electron in its "spin-up in the z direction" state, where $|\uparrow_z\rangle$ denotes this state. If we prepare the state in this way, we know for certain that the state is indeed $|\uparrow_z\rangle$ and no other state. Thus, we do not gain any information, or equivalently, there is no removal of uncertainty if someone else tells us that the state is $|\uparrow_z\rangle$. We may say that this state has zero qubits of quantum information, where the term "qubit" now refers to a measure of the quantum information of a state.

In the quantum world, we also have the option of measuring this state in the x direction. The postulates of quantum theory, given in Chapter 3, predict that the state will then be $|\uparrow_x\rangle$ or $|\downarrow_x\rangle$ with equal probability after measuring in the x direction. One interpretation of this aspect of quantum theory is that the system does not have any definite state in the x direction, in fact there is maximal uncertainty about its x direction, if we know that the physical system has a definite z direction. This behavior is one manifestation of the Heisenberg uncertainty principle. So before performing the measurement, we have no knowledge of the resulting state and we gain one Shannon bit of information after learning the result of the measurement. If we use Shannon's notion of entropy and perform an x measurement, this classical measure loses some of its capability here to capture our knowledge of the state of the system. It is inadequate to capture our knowledge of the state because we actually prepared it ourselves and know with certainty that it is in the state $|\uparrow_z\rangle$. With these different notions of information gain, which one is the most appropriate for the quantum case?

It turns out that the first way of thinking is the one that is most useful for quantifying quantum information. If someone tells us the quantum state of a particular physical system and this state is indeed the true state, then we have complete knowledge of the state and thus do not learn more "qubits" of quantum information from this point onward. This line of thinking is perhaps similar in one sense to the classical world, but different from the classical world, in the sense of the case presented in the previous paragraph.

Now suppose that a friend, let us call him "Bob," randomly prepares quantum states as a probabilistic ensemble. Suppose Bob prepares $|\uparrow_z\rangle$ or $|\downarrow_z\rangle$ with equal probability. With only this probabilistic knowledge, we acquire one bit of information if Bob reveals which state he prepared. We could also perform a quantum measurement on the system to determine what state Bob prepared (we discuss quantum measurements in detail in Chapter 3). One reasonable measurement to perform is a measurement in the z direction. The result of the measurement determines which state Bob actually prepared because both states in the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

ensembles are states with definite z direction. The result of this measurement thus gives us one bit of information—the same amount that we would learn if Bob informed us which state he prepared. It seems that most of this logic is similar to the classical case—i.e., the result of the measurement only gave us one Shannon bit of information.

Another measurement to perform is a measurement in the x direction. If the actual state prepared is $|\uparrow_z\rangle$, then the quantum theory predicts that the state becomes $|\uparrow_x\rangle$ or $|\downarrow_x\rangle$ with equal probability. Similarly, if the actual state prepared is $|\downarrow_z\rangle$, then the quantum theory predicts that the state again becomes $|\uparrow_x\rangle$ or $|\downarrow_x\rangle$ with equal probability. Calculating probabilities, the resulting state is $|\uparrow_x\rangle$ with probability 1/2 and $|\downarrow_x\rangle$ with probability 1/2. So the Shannon bit content of learning the result is again one bit, but we arrived at this conclusion in a much different fashion from the scenario where we measured in the z direction. How can we quantify the *quantum information* of this ensemble? We claim for now that this ensemble contains one *qubit* of quantum information and this result derives from either the measurement in the z direction or the measurement in the x direction for this particular ensemble.

Let us consider one final example that perhaps gives more insight into how we might quantify quantum information. Suppose Bob prepares $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$ with equal probability. The first state is spin-up in the z direction and the second is spin-up in the x direction. If Bob reveals which state he prepared, then we learn one Shannon bit of information. But suppose now that we would like to learn the prepared state on our own, without the help of our friend Bob. One possibility is to perform a measurement in the z direction. If the state prepared is $|\uparrow_z\rangle$, then we learn this result with probability 1/2. But if the state prepared is $|\uparrow_x\rangle$, then the quantum theory predicts that the state becomes $|\uparrow_z\rangle$ or $|\downarrow_z\rangle$ with equal probability (while we learn what the new state is). Thus, quantum theory predicts that the act of measuring this ensemble inevitably disturbs the state some of the time. Also, there is no way that we can learn with certainty whether the prepared state is $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$. Using a measurement in the z direction, the resulting state is $|\uparrow_z\rangle$ with probability 3/4 and $|\downarrow_z\rangle$ with probability 1/4. We learn less than one Shannon bit of information from this ensemble because the probability distribution becomes skewed when we perform this particular measurement.

The probabilities resulting from the measurement in the z direction are the same that would result from an ensemble where Bob prepares $|\uparrow_z\rangle$ with probability 3/4 and $|\downarrow_z\rangle$ with probability 1/4 and we perform a measurement in the z direction. The actual Shannon entropy of the distribution (3/4, 1/4) is about 0.81 bits, confirming our intuition that we learn approximately less than one bit. A similar, symmetric analysis holds to show that we gain 0.81 bits of information when we perform a measurement in the x direction.

We have more knowledge of the system in question if we gain less information from performing measurements on it. In the quantum theory, we learn less about a system if we perform a measurement on it that does not disturb it too much. Is there a measurement that we can perform in which we learn the least amount of information? Recall that learning the least amount of information is ideal because it has the interpretation that we require fewer questions on average to learn the result of a random experiment. Indeed, it turns out that a measurement in the $x+z$ direction reveals the least amount of information. Avoiding details

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

for now, this measurement returns a state that we label $|\uparrow_{x+z}\rangle$ with probability $\cos^2(\pi/8)$ and a state $|\downarrow_{x+z}\rangle$ with probability $\sin^2(\pi/8)$. This measurement has the desirable effect that it causes the least amount of disturbance to the original states in the ensemble. The entropy of the distribution resulting from the measurement is about 0.6 bits and is less than the one bit that we learn if Bob reveals the state. The entropy 0.6 is also the least amount of information among all possible sharp measurements that we may perform on the ensemble. We claim that this ensemble contains 0.6 *qubits* of quantum information.

We can determine the ultimate compressibility of classical data with Shannon's source coding theorem (we overview this technique in the next chapter). Is there a similar way that we can determine the ultimate compressibility of quantum information? This question was one of the early and profitable ones for quantum Shannon theory and the answer is affirmative. The technique for quantum compression is called Schumacher compression, named after Benjamin Schumacher. Schumacher used ideas similar to that of Shannon—he created the notion of a quantum information source that emits random physical qubits, and he invoked the law of large numbers to show that there is a so-called *typical subspace* where most of the quantum information really resides. This line of thought is similar to that which we will discuss in the overview of data compression in the next chapter. The size of the typical subspace for most quantum information sources is exponentially smaller than the size of the space in which the emitted physical qubits resides. Thus, one can "quantum compress" the quantum information to this subspace without losing much. Schumacher's quantum source coding theorem then quantifies, in an operational sense, the amount of actual quantum information that the ensemble contains. The amount of actual quantum information corresponds to the number of qubits, in the informational sense, that the ensemble contains. It is this measure that is equivalent to the "optimal measurement" one that we suggested in the previous paragraph. We will study this idea in more detail later when we introduce the quantum theory and a rigorous notion of a quantum information source.

The techniques of quantum Shannon theory are the direct *quantum* analog of the techniques from classical information theory. We use the law of large numbers and the notion of the typical subspace, but we require generalizations of measures from the classical world to determine how "close" two different quantum states are. One measure, the *fidelity*, has the operational interpretation that it gives the probability that one quantum state would pass a test for being another. The *trace distance* is another distance measure that is perhaps more similar to a classical distance measure—its classical analog is a measure of the closeness of two probability distributions. The techniques in quantum Shannon theory also reside firmly in the quantum theory and have no true classical analog for some cases. Some of the techniques will seem similar to those in the classical world, but the answer to some of the fundamental questions in quantum Shannon theory are far different from some of the answers in the classical world. It is the purpose of this book to explore the answers to the fundamental questions of quantum Shannon theory, and we now begin to ask what kinds of tasks we can perform.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

1.2.3 Operational Tasks in Quantum Shannon Theory

Quantum Shannon theory has several resources that two parties can exploit in a quantum information processing task. Perhaps the most natural quantum resource is a *noiseless qubit channel*. We can think of this resource as some medium through which a physical qubit can travel without being affected by any noise. One example of a noiseless qubit channel could be the free space through which a photon travels, where it ideally does not interact with any other particles along the way to its destination.[7]

A *noiseless classical bit channel* is a special case of a noiseless qubit channel because we can always encode classical information into quantum states. For the example of a photon, we can say that horizontal polarization corresponds to a '0' and vertical polarization corresponds to a '1'. We refer to the dynamic resource of a noiseless classical bit channel as a *cbit*, in order to distinguish it from the noiseless qubit channel.

Perhaps the most intriguing resource that two parties can share is noiseless entanglement. Any entanglement resource is a *static resource* because it is one that they share. Examples of static resources in the classical world are an information source that we would like to compress or a common secret key that two parties may possess. We actually have a way of measuring entanglement that we discuss later on, and for this reason, we can say that a sender and receiver have bits of entanglement or *ebits*.

Entanglement turns out to be a useful resource in many quantum communication tasks. One example where it is useful is in the teleportation protocol, where a sender and receiver use one ebit and two classical bits to transmit one qubit faithfully. This protocol is an example of the extraordinary power of noiseless entanglement. The name "teleportation" is really appropriate for this protocol because the physical qubit vanishes from the sender's station and appears at the receiver's station after the receiver obtains the two transmitted classical bits. We will see later on that a noiseless qubit channel can generate the other two noiseless resources, but it is impossible for each of the other two noiseless resources to generate the noiseless qubit channel. In this sense, the noiseless qubit channel is the strongest of the three unit resources.

The first quantum information processing task that we have discussed is Schumacher compression. The goal of this task is to use as few noiseless qubit channels as possible in order to transmit the output of a quantum information source reliably. After we understand Schumacher compression in a technical sense, the main focus of this book is to determine what quantum information processing tasks a sender and receiver can accomplish with the use of a noisy quantum channel. The first and perhaps simplest task is to determine how much classical information a sender can transmit reliably to a receiver, by using a noisy quantum channel a large number of times. This task is known as HSW coding, named after its discoverers Holevo, Schumacher, and Westmoreland. The HSW coding theorem is one quantum generalization of Shannon's channel coding theorem (overviewed in the next chapter). We can also assume that a sender and receiver share some amount of noiseless entanglement prior to communication. They can then use this noiseless entanglement in addition to a large

[7] We should be careful to note here that this is not actually a perfect channel because even empty space can be noisy in quantum mechanics, but nevertheless, it is a simple physical example to imagine.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

number of uses of a noisy quantum channel. This task is known as *entanglement-assisted classical communication* over a noisy quantum channel. The capacity theorem corresponding to this task again highlights one of the marvelous features of entanglement. It shows that entanglement gives a boost to the amount of noiseless classical communication we can generate with a noisy quantum channel—the classical capacity is generally higher with entanglement assistance than without it.

Perhaps the most important theorem for quantum Shannon theory is the *quantum channel capacity theorem*. Any proof of a capacity theorem consists of two parts: one part establishes a lower bound on the capacity and the other part establishes an upper bound. If the two bounds coincide, then we have a characterization of the capacity in terms of these bounds. The lower bound on the quantum capacity is colloquially known as the LSD coding theorem,[8] and it gives a characterization of the highest rate at which a sender can transmit quantum information reliably over a noisy quantum channel so that a receiver can recover it perfectly. The rate is generally lower than the classical capacity because it is more difficult to keep quantum information intact. As we have said before, it is possible to encode classical information into quantum states, but this classical encoding is only a special case of a quantum state. In order to preserve quantum information, we have to be able to preserve arbitrary quantum states, not merely a classical encoding within a quantum state.

The pinnacle of this book is in Chapter 23 where we finally reach our study of the quantum capacity theorem. All efforts and technical developments in preceding chapters have this goal in mind.[9] Our first coding theorem in the dynamic setting is the HSW coding theorem. A rigorous study of this coding theorem lays an important foundation—an understanding of a coding structure for a noisy quantum channel. The method for the HSW coding theorem applies to the "entanglement-assisted classical capacity theorem," which is one building block for other protocols in quantum Shannon theory. We then build a more complex coding structure for sending private classical information over a noisy quantum channel. In *private coding*, we are concerned with coding in such a way that the intended receiver can learn the transmitted message perfectly, but a third party eavesdropper cannot learn anything about what the sender transmits to the intended receiver. This study of the private classical capacity may seem like a detour at first, but it is closely linked with our ultimate aim. The coding structure developed for sending private information proves to be indispensable for understanding the structure of a quantum code. There are strong connections between the goals of keeping classical information private and keeping quantum information coherent. In the private coding scenario, the goal is to avoid leaking any information to an eavesdropper so that she cannot learn anything about the transmission. In the quantum coding scenario, we can think of quantum noise as resulting from the environment learning about the transmitted quantum information and this act of learning disturbs the quantum information. This

[8] The LSD coding theorem does not refer to the synthetic crystalline compound, lysergic acid diethylamide (which one may potentially use as a hallucinogenic drug), but refers rather to Seth Lloyd [185], Peter Shor [227], and Igor Devetak [68], all of whom gave separate proofs of the lower bound on the quantum capacity with increasing standards of rigor.

[9] One goal of this book is to unravel the mathematical machinery behind Devetak's proof of the quantum channel coding theorem [68].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

effect is related to the information-disturbance trade-off that is fundamental in quantum information theory. If the environment learns something about the state being transmitted, there is inevitably some sort of noisy disturbance that affects the quantum state. Thus, we can see a correspondence between private coding and quantum coding. In quantum coding, the goal is to avoid leaking any information to the environment because the avoidance of such a leak implies that there is no disturbance to the transmitted state. So the role of the environment in quantum coding is similar to the role of the eavesdropper in private coding, and the goal in both scenarios is to decouple either the environment or eavesdropper from the picture. It is then no coincidence that private codes and quantum codes have a similar structure. In fact, we can say that the quantum code inherits its structure from that of the private code.[10]

We also consider "trade-off" problems in addition to discussing the quantum capacity theorem. Chapter 21 is another high point of the book, featuring a whole host of results that emerge by combining several of the ideas from previous chapters. The most appealing aspect of this chapter is that we can construct virtually all of the protocols in quantum Shannon theory from just one idea in Chapter 20. Also, Chapter 21 answers many practical questions concerning information transmission over noisy quantum channels. Some example questions are as follows:

- How much quantum and classical information can a noisy quantum channel transmit?
- An entanglement-assisted noisy quantum channel can transmit more classical information than an unassisted one, but how much entanglement is really necessary?
- Does noiseless classical communication help in transmitting quantum information reliably over a noisy quantum channel?
- How much entanglement can a noisy quantum channel generate when aided by classical communication?
- How much quantum information can a noisy quantum channel communicate when aided by entanglement?

These are examples of trade-off problems because they involve a noisy quantum channel and either the consumption or generation of a noiseless resource. For every combination of the generation or consumption of a noiseless resource, there is a corresponding coding theorem that states what rates are achievable (and in some cases optimal). Some of these trade-off questions admit interesting answers, but some of them do not. Our final aim in these trade-off questions is to determine the full triple trade-off solution where we study the optimal ways of combining all three unit resources (classical communication, quantum communication, and entanglement) with a noisy quantum channel.

[10]There are other methods of formulating quantum codes using random subspaces [227, 132, 134, 174], but we prefer the approach of Devetak because we learn about other aspects of quantum Shannon theory, such as the private capacity, along the way to proving the quantum capacity theorem.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The coding theorems for a noisy quantum channel are just as important (if not more important) as Shannon's classical coding theorems because they determine the ultimate capabilities of information processing in a world where the postulates of quantum theory apply. It is thought that quantum theory is the ultimate theory underpinning all physical phenomena and any theory of gravity will have to incorporate the quantum theory in some fashion. Thus, it is reasonable that we should be focusing our efforts now on a full Shannon theory of quantum information processing in order to determine the tasks that these systems can accomplish. In many physical situations, some of the assumptions of quantum Shannon theory may not be justified (such as an independent and identically distributed quantum channel), but nevertheless, it provides an ideal setting in which we can determine the capabilities of these physical systems.

1.2.4 History of Quantum Shannon Theory

We conclude this introductory chapter by giving a brief overview of the problems that researchers were thinking about that ultimately led to the development of quantum Shannon theory.

The 1970s—The first researchers in quantum information theory were concerned with transmitting classical data by optical means. They were ultimately led to a quantum formulation because they wanted to transmit classical information by means of a coherent laser. *Coherent states* are special quantum states that a coherent laser ideally emits. Glauber provided a full quantum-mechanical theory of coherent states in two seminal papers [108, 109], for which he shared the Nobel Prize in 2005 [110]. The first researchers of quantum information theory were Helstrom, Gordon, Stratonovich, and Holevo. Gordon first conjectured an important bound for our ability to access classical information from a quantum system [111] and Levitin stated it without proof [182]. Holevo later provided a proof that the bound holds [143, 142]. This important bound is now known as the Holevo bound, and it is useful in proving converse theorems (theorems concerning optimality) in quantum Shannon theory. The simplest version of the Holevo bound states that it is not possible to transmit more than one classical bit of information using a noiseless qubit channel—i.e., we get *one cbit per qubit*. Helstrom developed a full theory of quantum detection and quantum estimation and published a textbook that discusses this theory [139]. Fannes contributed a useful continuity property of the entropy that is also useful in proving converse theorems in quantum Shannon theory [91]. Fannes was not studying quantum information theory proper, but nevertheless, his contribution, now known as the Fannes' inequality, is an important one for the field. Wiesner also used the uncertainty principle to devise a notion of "quantum money" in 1970, but unfortunately, his work was not accepted upon its initial submission. This work was *way* ahead of its time, and it was only until much later that it was accepted [247]. Wiesner's ideas paved the way for the BB84 protocol for quantum key distribution.

The 1980s—The 1980s witnessed only a few advances in quantum information theory because just a handful of researchers thought about the possibilities of linking quantum theory with information-theoretic ideas. The Nobel-prize winning physicist Richard Feynman published an interesting 1982 article that was one of the first to discuss computing with

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

quantum-mechanical systems [94]. His interest was in using a quantum computer to simulate quantum-mechanical systems—he figured there should be a speed-up over a classical simulation if we instead use a quantum system to simulate another. This work is less quantum Shannon theory than it is quantum computing, but it is still a landmark because Feynman began to think about exploiting the actual quantum information in a physical system, rather than just using quantum systems to process classical information as the researchers in the 1970s suggested.

Wootters and Zurek produced one of the simplest, yet most profound, results that is crucial to quantum information science [265] (Dieks also proved this result in the same year [78]). They proved the *no-cloning theorem*, showing that the postulates of the quantum theory imply the impossibility of universally cloning quantum states. Given an arbitrary unknown quantum state, it is impossible to build a device that can copy this state. This result has deep implications for the processing of quantum information and shows a strong divide between information processing in the quantum world and that in the classical world. We will prove this theorem in Chapter 3 and use it time and again in our reasoning. The history of the no-cloning theorem is one of the more interesting "sociology of science" stories that you may come across. The story goes that Nick Herbert submitted a paper to *Foundations of Physics* with a proposal for faster-than-light communication using entanglement. Asher Peres was the referee [203], and he knew that something had to be wrong with the proposal because it allowed for superluminal communication, yet he could not put his finger on what the problem might be (he also figured that Herbert knew his proposal was flawed). Nevertheless, Peres recommended the paper for publication [140] because he figured it would stimulate wide interest in the topic. Not much later, Wootters and Zurek published their paper, and since then, there have been thousands of follow-up results on the no-cloning theorem [213].

The work of Wiesner on conjugate coding inspired an IBM physicist named Charles Bennett. In 1984, Bennett and Brassard published a groundbreaking paper that detailed the first quantum communication protocol: the BB84 protocol [22]. This protocol shows how a sender and a receiver can exploit a quantum channel to establish a secret key. The security of this protocol relies on the uncertainty principle. If any eavesdropper tries to learn about the random quantum data that they use to establish the secret key, this act of learning inevitably disturbs the transmitted quantum data and the two parties can discover this disturbance by noticing the change in the statistics of random sample data. The secret key generation capacity of a noisy quantum channel is inextricably linked to the BB84 protocol, and we study this capacity problem in detail when we study the ability of quantum channels to communicate private information. Interestingly, the physics community largely ignored the BB84 paper when Bennett and Brassard first published it, likely because they presented it at an engineering conference and the merging of physics and information had not yet taken effect.

The 1990s—The 1990s were a time of much increased activity in quantum information science, perhaps some of the most exciting years with many seminal results. One of the first major results was from Ekert. He published a different way for performing quantum key distribution, this time relying on the strong correlations of entanglement [88]. He was

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

unaware of the BB84 protocol when he was working on his entanglement-based quantum key distribution. The physics community embraced this result and shortly later, Ekert and Bennett and Brassard became aware of each other's respective works [24]. Bennett, Brassard, and Mermin later showed a sense in which these two seemingly different schemes are equivalent [25]. Bennett later developed the B92 protocol for quantum key distribution using any two non-orthogonal quantum states [18].

Two of the most profound results that later impacted quantum Shannon theory appeared in the early 1990s. First, Bennett and Wiesner devised the super-dense coding protocol [35]. This protocol consumes one noiseless ebit of entanglement and one noiseless qubit channel to simulate two noiseless classical bit channels. Let us compare this result to that of Holevo. Holevo's bound states that we can only send one classical bit per qubit, but the super-dense coding protocol states that we can double this rate if we consume entanglement as well. Thus, entanglement is the enabler in this protocol that boosts the classical rate beyond that possible with a noiseless qubit channel alone. The next year, Bennett and some other coauthors reversed the operations in the super-dense coding protocol to devise a protocol that has more profound implications. They devised the *teleportation protocol* [23]—this protocol consumes two classical bit channels and one ebit to transmit a qubit from a sender to receiver. Right now, without any technical development yet, it may be unclear how the qubit gets from the sender to receiver. The original authors described it as the "disembodied transport of a quantum state." Suffice it for now to say that it is the unique properties of entanglement (in particular, the ebit) that enable this disembodied transport to occur. Yet again, it is entanglement that is the resource that enables this protocol, but let us be careful not to overstate the role of entanglement. Entanglement alone cannot do much. These protocols show that it is the unique combination of entanglement and quantum communication or entanglement and classical communication that yields these results. These two noiseless protocols are cornerstones of quantum Shannon theory, originally suggesting that there are interesting ways of combining the resources of classical communication, quantum communication, and entanglement to formulate uniquely quantum protocols and leading the way to more exotic protocols that combine the different noiseless resources with noisy resources. Simple questions concerning these protocols lead to quantum Shannon-theoretic protocols. In super-dense coding, how much classical information can Alice send if the quantum channel becomes noisy? What if the entanglement is noisy? In teleportation, how much quantum information can Alice send if the classical channel is noisy? What if the entanglement is noisy? Researchers addressed these questions quite a bit after the original super-dense coding and teleportation protocols were available, and we address these important questions in this book.

The year 1994 was a landmark for quantum information science. Shor published his algorithm that factors a number in polynomial time [223]—this algorithm gives an exponential speedup over the best known classical algorithm. We cannot overstate the importance of this algorithm for the field. Its major application is to break RSA encryption [208] because the security of that encryption algorithm relies on the computational difficulty of factoring a large number. This breakthrough generated wide interest in the idea of a quantum computer

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and started the quest to build one and study its abilities.

Initially, much skepticism met the idea of building a practical quantum computer [181, 241]. Some experts thought that it would be impossible to overcome errors that inevitably occur during quantum interactions, due to the coupling of a quantum system with its environment. Shor met this challenge by devising the first quantum error-correcting code [224] and a scheme for fault-tolerant quantum computation [225]. His paper on quantum error correction is the one most relevant for quantum Shannon theory. At the end of this paper, he posed the idea of the quantum capacity of a noisy quantum channel as the highest rate at which a sender and receiver can maintain the fidelity of a quantum state through a large number of uses of the noisy channel. This open problem set the main task for researchers interested in quantum Shannon theory. A flurry of theoretical activity then ensued in quantum error correction [54, 235, 180, 112, 113, 52, 53] and fault-tolerant quantum computation [6, 173, 206, 175]. These two areas are now important subfields within quantum information science, but we do not focus on them in any detail in this book.

Schumacher published a critical paper in 1995 as well [216] (we discussed some of his contributions in the previous section). This paper gave the first informational notion of a qubit, and it even established the now ubiquitous term "qubit." He proved the quantum analog of Shannon's source coding theorem, giving the ultimate compressibility of quantum information. He used the notion of a typical subspace as an analogy of Shannon's typical set. This notion of a typical subspace proves to be one of the most crucial ideas for constructing codes in quantum Shannon theory, just as the notion of a typical set is so crucial for Shannon's information theory.

Not much later, several researchers began investigating the capacity of a noisy quantum channel for sending classical information [126]. Holevo [144] and Schumacher and Westmoreland [219] independently proved that the Holevo information of a quantum channel is an achievable rate for classical communication over it. They appealed to Schumacher's notion of a typical subspace and constructed channel codes for sending classical information. The proof looks somewhat similar to the proof of Shannon's channel coding theorem (discussed in the next chapter) after taking a few steps away from it. The proof of the converse theorem proceeds somewhat analogously to that of Shannon's theorem, with the exception that one of the steps uses Holevo's bound from 1973. It is perhaps somewhat surprising that it took over thirty years between the appearance of the proof of Holevo's bound (the main step in the converse proof) and the appearance of a direct coding theorem for sending classical information.

The quantum capacity theorem is perhaps the fundamental theorem of quantum Shannon theory. Initial work by several researchers provided some insight into the quantum capacity theorem [26, 30, 29, 220], and a series of papers by Barnum, Knill, Nielsen, and Schumacher established an upper bound on the quantum capacity [217, 218, 16, 15]. For the lower bound, Lloyd was the first to construct an idea for a proof, but it turns out that his proof was more of a heuristic proof [185]. Shor then followed with another proof of the lower bound [227], and some of Shor's ideas appeared much later in a full publication [134]. Devetak [68] and Cai, Winter, and Yeung [51] independently solved the private capacity theorem

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

at approximately the same time (with the publication of the CWY paper appearing a year after Devetak's arXiv post). Devetak took the proof of the private capacity theorem a step further and showed how to apply its techniques to construct a quantum code that achieves a good lower bound on the quantum capacity, while also providing an alternate, cleaner proof of the converse theorem [68]. It is Devetak's technique that we mainly explore in this book because it provides some insight into the coding structure (though, we also explore a different technique via the entanglement-assisted classical capacity theorem).

The 2000s—In recent years, we have had many advancements in quantum Shannon theory (technically some of the above contributions were in the 2000s, but we did not want to break the continuity of the history of the quantum capacity theorem). One major result was the proof of the entanglement-assisted classical capacity theorem—it is the noisy version of the super-dense coding protocol where the quantum channel is noisy [33, 34, 146]. This theorem assumes that Alice and Bob share unlimited entanglement and they exploit the entanglement and the noisy quantum channel to send classical information.

A few fantastic results have arisen in recent years. Horodecki, Oppenheim, and Winter showed the existence of a state-merging protocol [148, 149]. This protocol gives the minimum rate at which Alice and Bob consume noiseless qubit channels in order for Alice to send her part of a quantum state to Bob. This rate is the conditional quantum entropy—the protocol thus gives an operational interpretation to this entropic quantity. What was most fascinating about this result is that the conditional quantum entropy can be negative in quantum Shannon theory. Prior to their work, no one really understood what it meant for the conditional quantum entropy to become negative [246, 154, 55], but this state merging result gave a good operational interpretation. A negative rate implies that Alice and Bob gain the ability for future quantum communication, instead of consuming quantum communication as when the rate is positive.

Another fantastic result came from Smith and Yard [234]. Suppose we have two noisy quantum channels and each of them individually has zero capacity to transmit quantum information. One would expect intuitively that the "joint quantum capacity" (when using them together) would also have zero ability to transmit quantum information. But this result is not generally the case in the quantum world. It is possible for some particular noisy quantum channels with no individual quantum capacity to have a non-zero joint quantum capacity. It is not clear yet how we might practically take advantage of such a "superactivation" effect, but the result is nonetheless fascinating, counterintuitive, and not yet fully understood.

The latter part of this decade has seen the unification of quantum Shannon theory. The resource inequality framework was the first step because it unified many previously known results into one formalism [71, 70]. Devetak, Harrow, and Winter provided a family tree for quantum Shannon theory and showed how to relate the different protocols in the tree to one another. We will go into the theory of resource inequalities in some detail throughout this book because it provides a tremendous conceptual simplification when considering coding theorems in quantum Shannon theory. In fact, the last chapter of this book contains a concise summary of many of the major quantum Shannon-theoretic protocols in the language of

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

resource inequalities. Abeyesinghe, Devetak, Hayden, and Winter published a work showing a sense in which the mother protocol of the family tree can generate the father protocol [3]. We have seen unification efforts in the form of triple trade-off coding theorems [4, 159, 160]. These theorems give the optimal combination of classical communication, quantum communication, entanglement, and an asymptotic noisy resource for achieving a variety of quantum information processing tasks.

We have also witnessed the emergence of a study of network quantum Shannon theory. Some authors have tackled the quantum broadcasting paradigm [270, 84, 118, 119], where one sender transmits to multiple receivers. A multiple-access quantum channel has many senders and one receiver. Some of the same authors (and others) have tackled multiple-access communication [257, 269, 266, 272, 268, 156, 59]. This network quantum Shannon theory should become increasingly important as we get closer to the ultimate goal of a quantum Internet.

Quantum Shannon theory has now established itself as an important and distinct field of study. The next few chapters discuss the concepts that will prepare us for tackling some of the major results in quantum Shannon theory.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 2

Classical Shannon Theory

We cannot overstate the importance of Shannon's contribution to modern science. His introduction of the field of information theory and his solutions to its two main theorems demonstrate that his ideas on communication were far beyond the other prevailing ideas in this domain around 1948.

In this chapter, our aim is to discuss Shannon's two main contributions in a descriptive fashion. The goal of this high-level discussion is to build up the intuition for the problem domain of information theory and to understand the main concepts before we delve into the analogous quantum information-theoretic ideas. We avoid going into deep technical detail in this chapter, leaving such details for later chapters where we formally prove both classical and quantum Shannon-theoretic coding theorems. We do use some mathematics from probability theory, namely, the law of large numbers.

We will be delving into the technical details of this chapter's material in later chapters (specifically, Chapters 10, 12, and 13). Once you have reached later chapters that develop some more technical details, it might be helpful to turn back to this chapter to get an overall flavor for the motivation of the development.

2.1 Data Compression

We first discuss the problem of data compression. Those who are familiar with the Internet have used several popular data formats as JPEG, MPEG, ZIP, GIF, etc. All of these file formats have corresponding algorithms for compressing the output of an information source. A first glance at the compression problem may lead one to believe that it is possible to compress the output of the information source to an arbitrarily small size, but Shannon proved that arbitrarily small compression is not possible. This result is the content of Shannon's first noiseless coding theorem.

2.1.1 An Example of Data Compression

We begin with a simple example that illustrates the concept of an information source. We then develop a scheme for coding this source so that it requires fewer bits to represent its output faithfully.

Suppose that Alice is a sender and Bob is a receiver. Suppose further that a noiseless bit channel connects Alice to Bob—a noiseless bit channel is one that transmits information perfectly from sender to receiver, e.g., Bob receives '0' if Alice transmits '0' and Bob receives '1' if Alice transmits '1'. Alice and Bob would like to minimize the number of times that they use this noiseless channel because it is expensive to use it.

Alice would like to use the noiseless channel to communicate information to Bob. Suppose that an information source randomly chooses from four symbols $\{a, b, c, d\}$ and selects them with a skewed probability distribution:

$$\Pr\{a\} = 1/2, \tag{2.1}$$
$$\Pr\{b\} = 1/8, \tag{2.2}$$
$$\Pr\{c\} = 1/4, \tag{2.3}$$
$$\Pr\{d\} = 1/8. \tag{2.4}$$

So it is clear that the symbol a is the most likely one, c the next likely, and both b and d are least likely. We make the additional assumption that the information source chooses each symbol independently of all previous ones and chooses each with the same probability distribution above. After the information source makes a selection, it gives the symbol to Alice for coding.

A noiseless bit channel only accepts bits—it does not accept the symbols a, b, c, d as input. So, Alice has to encode her information into bits. Alice could use the following coding scheme:

$$a \to 00, \quad b \to 01, \quad c \to 10, \quad d \to 11, \tag{2.5}$$

where each binary representation of a letter is a *codeword*. How do we measure the performance of a particular coding scheme? The expected length of a codeword is one way to measure performance. For the above example, the expected length is equal to two bits. This measure reveals a problem with the above scheme—the scheme does not take advantage of the skewed nature of the distribution of the information source because each codeword is the same length.

One might instead consider a scheme that uses shorter codewords for symbols that are more likely and longer codewords for symbols that are less likely.[1] Then the expected length

[1] Such coding schemes are common. Samuel F. B. Morse employed this idea in his popular Morse code. Also, in the movie *The Diving Bell and the Butterfly*, a writer becomes paralyzed with "locked-in" syndrome so that he can only blink his left eye. An assistant then develops a "blinking code" where she reads a list of letters in French, beginning with the most commonly used letter and ending with the least commonly used letter. The writer blinks when she says the letter he wishes and they finish an entire book with this coding scheme.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

of a codeword should be shorter than the expected length of a codeword in the above scheme. The following coding scheme gives an improvement in the expected length of a codeword:

$$a \to 0, \quad b \to 110, \quad c \to 10, \quad d \to 111. \tag{2.6}$$

The above scheme has the advantage that any coded sequence is uniquely decodable. For example, suppose that Bob obtains the following sequence:

$$0011010111010100010. \tag{2.7}$$

Bob can parse the above sequence as

$$0\ 0\ 110\ 10\ 111\ 0\ 10\ 10\ 0\ 0\ 10, \tag{2.8}$$

and determine that Alice transmitted the message

$$aabcdaccaac. \tag{2.9}$$

We can calculate the expected length of this coding scheme as follows:

$$\frac{1}{2}(1) + \frac{1}{8}(3) + \frac{1}{4}(2) + \frac{1}{8}(3) = \frac{7}{4}. \tag{2.10}$$

This scheme is thus more efficient because its expected length is 7/4 bits as opposed to two bits. It is a *variable-length code* because the number of bits in each codeword depends on the source symbol.

2.1.2 A Measure of Information

The above scheme suggests a way to measure information. Consider the probability distribution in (2.4). Would we be more surprised to learn that the information source produced the symbol a or to learn that it produced the symbol d? The answer is d because the source is less likely to produce it. One measure of the surprise of symbol $x \in \{a, b, c, d\}$ is

$$i(x) \equiv \log\left(\frac{1}{p(x)}\right) = -\log(p(x)), \tag{2.11}$$

where the logarithm is base two—this convention implies the units of this measure are bits. This measure of surprise has the desirable property that it is higher for lower probability events and lower for higher probability events. Here, we take after Shannon, and we name $i(x)$ the *information content* of the symbol x. Observe that the length of each codeword in the coding scheme in (2.6) is equal to the information content of its corresponding symbol.

The information content has another desirable property called *additivity*. Suppose that the information source produces two symbols x_1 and x_2. The probability for this event is

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$p(x_1, x_2)$ and the joint distribution factors as $p(x_1)p(x_2)$ if we assume the source is *memoryless*—it produces each symbol independently. The information content of the two symbols x_1 and x_2 is additive because

$$\begin{aligned} i(x_1, x_2) &= -\log(p(x_1, x_2)) && (2.12)\\ &= -\log(p(x_1)p(x_2)) && (2.13)\\ &= -\log(p(x_1)) - \log(p(x_1)) && (2.14)\\ &= i(x_1) + i(x_2). && (2.15) \end{aligned}$$

In general, additivity is a desirable property for any information measure. We will return to the issue of additivity in many different contexts in this book (especially in Chapter 12).

The expected information content of the information source is

$$\sum_x p(x)i(x) = -\sum_x p(x)\log(p(x)). \tag{2.16}$$

The above quantity is so important in information theory that we give it a name: the *entropy* of the information source. The reason for its importance is that the entropy and variations of it appear as the answer to many questions in information theory. For example, in the above coding scheme, the expected length of a codeword is the entropy of the information source because

$$\begin{aligned} &-\frac{1}{2}\log\frac{1}{2} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8}\\ &\quad= \frac{1}{2}(1) + \frac{1}{8}(3) + \frac{1}{4}(2) + \frac{1}{8}(3) && (2.17)\\ &\quad= \frac{7}{4}. && (2.18) \end{aligned}$$

It is no coincidence that we chose the particular coding scheme in (2.6). The effectiveness of the scheme in this example is related to the structure of the information source—the number of symbols is a power of two and the probability of each symbol is the reciprocal of a power of two.

2.1.3 Shannon's Source Coding Theorem

The next question to ask is whether there is any other scheme that can achieve a better compression rate than the scheme in (2.6). This question is the one that Shannon asked in his first coding theorem. To answer this question, we consider a more general information source and introduce a notion of Shannon, the idea of the *set of typical sequences*.

We can represent a more general information source with a random variable X whose realizations x are *letters* in an *alphabet* $\mathcal{X}$. Let $p_X(x)$ be the probability mass function associated with random variable X, so that the probability of realization x is $p_X(x)$. Let $H(X)$ denote the entropy of the information source:

$$H(X) \equiv -\sum_{x\in\mathcal{X}} p_X(x)\log(p_X(x)). \tag{2.19}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The entropy $H(X)$ is also the entropy of the random variable X. Another way of writing it is $H(p)$, but we use the more common notation $H(X)$ throughout this book.

The information content $i(X)$ of random variable X is

$$i(X) \equiv -\log(p_X(X)), \tag{2.20}$$

and is itself a random variable. There is nothing wrong mathematically here with having random variable X as the argument to the density function p_X, though this expression may seem self-referential at a first glance. This way of thinking turns out to be useful later. Again, the expected information content of X is equal to the entropy:

$$\mathbb{E}_X\{-\log(p_X(X))\} = H(X). \tag{2.21}$$

Exercise 2.1.1 Show that the entropy of a uniform random variable is equal to $\log|\mathcal{X}|$ where $|\mathcal{X}|$ is the size of the variable's alphabet.

We now turn to source coding the above information source. We *could* associate a binary codeword for each symbol x as we did in the scheme in (2.6). But this scheme may lose some efficiency if the size of our alphabet is not a power of two or if the probabilities are not a reciprocal of a power of two as they are in our nice example. Shannon's breakthrough idea was to let the source emit a large number of realizations and then code the emitted data as a large block, instead of coding each symbol as the above example does. This technique is called *block coding*. Shannon's other insight was to allow for a slight error in the compression scheme, but show that this error vanishes as the block size becomes arbitrarily large. To make the block coding scheme more clear, Shannon suggests to let the source emit the following sequence:

$$x^n \equiv x_1 x_2 \cdots x_n, \tag{2.22}$$

where n is a large number that denotes the size of the block of emitted data and x_i, for all $i = 1, \ldots, n$, denotes the i^{th} emitted symbol. Let X^n denote the random variable associated with the sequence x^n, and let X_i be the random variable for the i^{th} symbol x_i. Figure 2.1 depicts Shannon's idea for a classical source code.

The most important assumption for this information source is that it is independent and identically distributed (IID). The IID assumption means that each random variable X_i has the same distribution as random variable X, and we use the index i merely to track to which symbol x_i the random variable X_i corresponds. Under the IID assumption, the probability of any given emitted sequence x^n factors as

$$\begin{aligned} p_{X^n}(x^n) &= p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) && (2.23)\\ &= p_{X_1}(x_1) p_{X_2}(x_2) \cdots p_{X_n}(x_n) && (2.24)\\ &= p_X(x_1) p_X(x_2) \cdots p_X(x_n) && (2.25)\\ &= \prod_{i=1}^{n} p_X(x_i). && (2.26) \end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

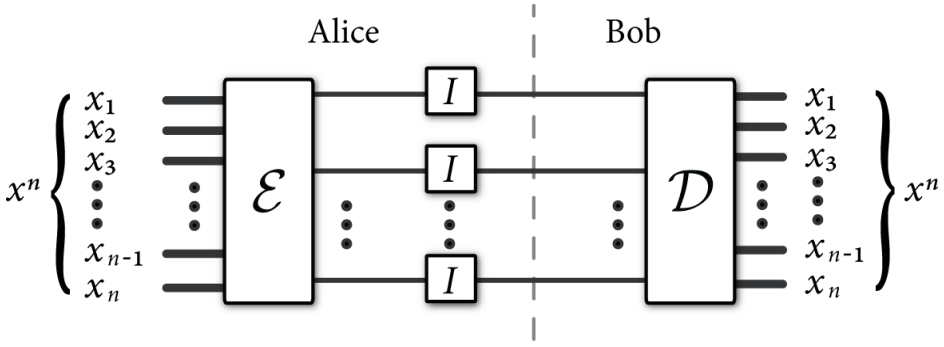

Figure 2.1: The above figure depicts Shannon's idea for a classical source code. The information source emits a long sequence x^n to Alice. She encodes this sequence as a block with an encoder $\mathcal{E}$ and produces a codeword whose length is less than that of the original sequence x^n (indicated by fewer lines coming out of the encoder $\mathcal{E}$). She transmits the codeword over noiseless bit channels (each indicated by "I" which stands for the identity bit channel) and Bob receives it. Bob decodes the transmitted codeword with a decoder $\mathcal{D}$ and produces the original sequence that Alice transmitted, only if their chosen code is good, in the sense that the code has a small probability of error.

The above rule from probability theory results in a remarkable simplification of the mathematics. Suppose that we now label the letters in the alphabet $\mathcal{X}$ as $a_1, \ldots, a_{|\mathcal{X}|}$ in order to distinguish the letters from the realizations. Let $N(a_i|x^n)$ denote the number of occurrences of the letter a_i in the sequence x^n (where $i = 1, \ldots, |\mathcal{X}|$). As an example, consider the sequence in (2.9). The quantities $N(a_i|x^n)$ for this example are

$$N(a|x^n) = 5, \tag{2.27}$$
$$N(b|x^n) = 1, \tag{2.28}$$
$$N(c|x^n) = 4, \tag{2.29}$$
$$N(d|x^n) = 1. \tag{2.30}$$

We can rewrite the result in (2.26) as

$$p_{X^n}(x^n) = \prod_{i=1}^{n} p_X(x_i) = \prod_{i=1}^{|\mathcal{X}|} p_X(a_i)^{N(a_i|x^n)} \tag{2.31}$$

Keep in mind that we are allowing the length n of the emitted sequence to be extremely large so that it is much larger than the alphabet size $|\mathcal{X}|$:

$$n \gg |\mathcal{X}|. \tag{2.32}$$

The formula on the right in (2.31) is much simpler than the the formula in (2.26) because it has fewer iterations of multiplications. There is a sense in which the IID assumption allows us to permute the sequence x^n as

$$x^n \rightarrow \underbrace{a_1 \cdots a_1}_{N(a_1|x^n)} \underbrace{a_2 \cdots a_2}_{N(a_2|x^n)} \cdots \underbrace{a_{|\mathcal{X}|} \cdots a_{|\mathcal{X}|}}_{N\left(a_{|\mathcal{X}|}|x^n\right)}, \tag{2.33}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

because the probability calculation is invariant under this permutation. We introduce the above way of thinking right now because it turns out to be useful later when we develop some ideas in quantum Shannon theory (specifically in Section 13.9. Thus, the formula on the right in (2.31) characterizes the probability of any given sequence x^n.

The above discussion applies to a particular sequence x^n that the information source emits. Now, we would like to analyze the behavior of a *random sequence* X^n that the source emits, and this distinction between the realization x^n and the random variable X^n is important. In particular, let us consider the sample average of the information content of the random sequence X^n (divide the information content of X^n by n to get the sample average):

$$-\frac{1}{n}\log(p_{X^n}(X^n)). \tag{2.34}$$

It may seem strange at first glance that X^n, the argument of the probability mass function p_{X^n} is itself a random variable, but this type of expression is perfectly well defined mathematically. (This self-referencing type of expression is similar to (2.20), which we used to calculate the entropy.) For reasons that will become clear shortly, we call the above quantity *the sample entropy* of the random sequence X^n.

Suppose now that we use the function $N(a_i|\bullet)$ to calculate the number of appearances of the letter a_i in the random sequence X^n. We write the desired quantity as $N(a_i|X^n)$ and note that it is also a random variable, whose random nature derives from that of X^n. We can reduce the expression in (2.34) to the following one with some algebra and the result in (2.31):

$$-\frac{1}{n}\log(p_{X^n}(X^n)) = -\frac{1}{n}\log\left(\prod_{i=1}^{|\mathcal{X}|} p_X(a_i)^{N(a_i|X^n)}\right) \tag{2.35}$$

$$= -\frac{1}{n}\sum_{i=1}^{|\mathcal{X}|}\log\left(p_X(a_i)^{N(a_i|X^n)}\right) \tag{2.36}$$

$$= -\sum_{i=1}^{|\mathcal{X}|}\frac{N(a_i|X^n)}{n}\log(p_X(a_i)). \tag{2.37}$$

We stress again that the above quantity is random.

Is there any way that we can determine the behavior of the above sample entropy when n becomes large? Probability theory gives us a way. The expression $N(a_i|X^n)/n$ represents an empirical distribution for the letters a_i in the alphabet $\mathcal{X}$. As n becomes large, one form of the law of large numbers states that it is overwhelmingly likely that a random sequence has its empirical distribution $N(a_i|X^n)/n$ close to the true distribution $p_X(a_i)$, and conversely, it is highly unlikely that a random sequence does not satisfy this property. Thus, a random emitted sequence X^n is highly likely to satisfy the following condition for all $\delta > 0$ as n

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

becomes large:

$$\lim_{n\to\infty} \Pr\left\{ \left| -\frac{1}{n}\log(p_{X^n}(X^n)) - \sum_{i=1}^{|\mathcal{X}|} p_X(a_i)\log\left(\frac{1}{p_X(a_i)}\right)\right| \leq \delta \right\} = 1. \tag{2.38}$$

The quantity $-\sum_{i=1}^{|\mathcal{X}|} p_X(a_i)\log(p_X(a_i))$ is none other than the entropy $H(X)$ so that the above expression is equivalent to the following one for all $\delta > 0$:

$$\lim_{n\to\infty} \Pr\left\{ \left| -\frac{1}{n}\log(p_{X^n}(X^n)) - H(X)\right| \leq \delta \right\} = 1. \tag{2.39}$$

Another way of stating this property is as follows:

> It is highly likely that the information source emits a sequence whose sample entropy is close to the true entropy, and conversely, it is highly unlikely that the information source emits a sequence that does not satisfy this property.[2]

Now we consider a particular realization x^n of the random sequence X^n. We name a particular sequence x^n a *typical sequence* if its sample entropy is close to the true entropy $H(X)$ and the set of all typical sequences is the *typical set*. Fortunately for data compression, the set of typical sequences is not too large. In Chapter 13 on typical sequences, we prove that the size of this set is much smaller than the set of all sequences. We accept it for now (and prove later) that the size of the typical set is $\approx 2^{nH(X)}$, whereas the size of the set of all sequences is equal to $|\mathcal{X}|^n$. We can rewrite the size of the set of all sequences as

$$|\mathcal{X}|^n = 2^{n\log|\mathcal{X}|}. \tag{2.40}$$

Comparing the size of the typical set to the size of the set of all sequences, the typical set is exponentially smaller than the set of all sequences whenever the random variable is not equal to the uniform random variable. Figure 2.2 illustrates this concept. We summarize these two crucial properties of the typical set and give another that we prove later:

Property 2.1.1 (Unit Probability) The probability that an emitted sequence is typical approaches unity as n becomes large. Another way of stating this property is that the typical set has almost all of the probability.

Property 2.1.2 (Exponentially Small Cardinality) The size of the typical set is $2^{nH(X)}$ and is exponentially smaller than the size $2^{n\log|\mathcal{X}|}$ of the set of all sequences whenever random variable X is not uniform.

[2]Do not fall into the trap of thinking "The possible sequences that the source emits are typical sequences." That line of reasoning is quantitatively far from the truth. In fact, what we can show is much different because the set of typical sequences is much smaller than the set of all possible sequences.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

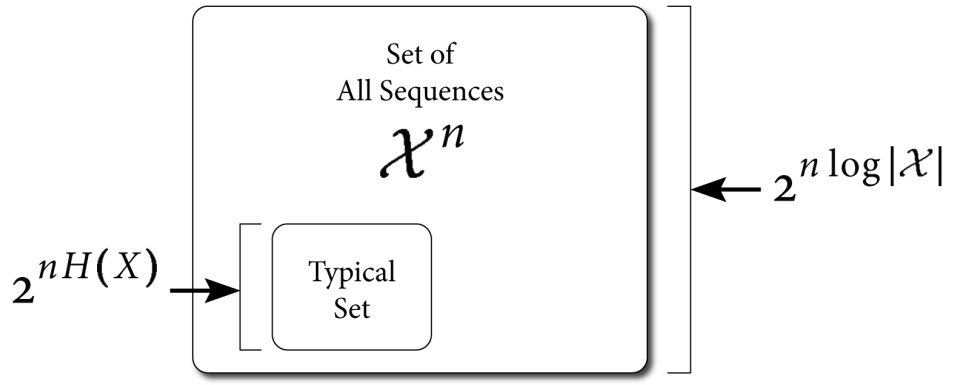

Figure 2.2: The above figure indicates that the typical set is much smaller (exponentially smaller) than the set of all sequences. The typical set is the same size as the set of all sequences only when the entropy $H(X)$ of the random variable X is equal to $\log|\mathcal{X}|$—implying that the distribution of random variable X is uniform.

Property 2.1.3 (Equipartition) The probability of a *particular* typical sequence is roughly uniform $\approx 2^{-nH(X)}$. (The probability $2^{-nH(X)}$ is easy to calculate if we accept that the typical set has all of the probability, its size is $2^{nH(X)}$, and the distribution over typical sequences is uniform.)

These three properties together are collectively known as the *asymptotic equipartition theorem*. The word "asymptotic" applies because the theorem exploits the asymptotic limit when n is large and the word "equipartition" refers to the third property above.

With the above notions of a typical set under our belt, a strategy for compressing information should now be clear. The strategy is to compress only the typical sequences that the source emits. We simply need to establish an invertible encoding function that maps from the set of typical sequences (size $2^{nH(X)}$) to the set of all binary strings of length $nH(X)$ (this set also has size $2^{nH(X)}$). If the source emits an atypical sequence, we declare an error. This coding scheme is reliable in the asymptotic limit because the probability of an error event vanishes as n becomes large, thanks to the unit probability property in the asymptotic equipartition theorem. We measure the rate of this block coding scheme as follows:

$$\text{compression rate} \equiv \frac{\#\text{ of noiseless channel bits}}{\#\text{ of source symbols}}. \tag{2.41}$$

For the case of Shannon compression, the number of noiseless channel bits is equal to $nH(X)$ and the number of source symbols is equal to n. Thus, the rate is $H(X)$ and this protocol gives an *operational interpretation* to the Shannon entropy $H(X)$ because it appears as the rate of data compression.

One may then wonder whether this rate of data compression is the best that we can do—whether this rate is optimal (we could achieve a lower rate of compression if it were not optimal). In fact, the above rate is the optimal rate at which we can compress information. We hold off on a formal proof of optimality for now and delay it until we reach Chapter 17.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above discussion highlights the common approach in information theory for establishing a coding theorem. Proving a coding theorem has two parts—traditionally called the *direct coding theorem* and the *converse theorem*. First, we give a coding scheme that can achieve a given rate for an information processing task. This first part includes a direct construction of a coding scheme, hence the name *direct coding theorem*. The formal statement of the direct coding theorem for the above task is

> "If the rate of compression is greater than the entropy of the source, then there exists a coding scheme that can achieve lossless data compression in the sense that it is possible to make the probability of error for incorrectly decoding arbitrarily small."

The second task is to prove that the rate from the direct coding theorem is optimal—that we cannot do any better than the suggested rate. We traditionally call this part the converse theorem because it formally corresponds to the converse of the above statement:

> "If there exists a coding scheme that can achieve lossless data compression with arbitrarily small probability of decoding error, then the rate of compression is greater than the entropy of the source."

The techniques used in proving each part of the coding theorem are completely different. For most coding theorems in information theory, we can prove the direct coding theorem by appealing to the ideas of typical sequences and large block sizes. That this technique gives a good coding scheme is directly related to the asymptotic equipartition theorem properties that govern the behavior of random sequences of data as the length of the sequence becomes large. The proof of a converse theorem relies on information inequalities that give tight bounds on the entropic quantities appearing in the coding constructions. We spend some time with information inequalities in Chapter 10 to build up our ability to prove converse theorems.

Sometimes, in the course of proving a direct coding theorem, one may think to have found the optimal rate for a given information processing task. Without a matching converse theorem, it is not generally clear that the suggested rate is optimal. So, always prove converse theorems!

2.2 Channel Capacity

The next issue that we overview is the transmission of information over a noisy classical channel. We begin with a standard example—transmitting a single bit of information over a noisy bit flip channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

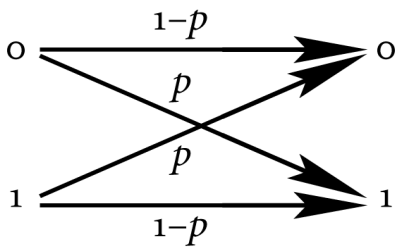

Figure 2.3: The above figure depicts the action of the bit-flip channel. It preserves the input bit with probability $1 - p$ and flips it with probability p.

2.2.1 An Example of an Error Correction Code

We again have our protagonists, Alice and Bob, as respective sender and receiver. This time, though, we assume that a noisy classical channel connects them, so that information transfer is not reliable. Alice and Bob realize that a noisy channel is not as expensive as a noiseless one, but it still is expensive for them to use. For this reason, they would like to maximize the amount of information that Alice can communicate reliably to Bob, where reliable communication implies that there is zero probability of error for every transmission.

The simplest example of a noisy classical channel is a bit-flip channel, with the technical name *binary symmetric channel*. This channel flips the input bit with probability p and leaves it unchanged with probability $1 - p$. Figure 2.3 depicts the action of the bit-flip channel. The channel behaves independently from one use to the next and behaves in the same random way as described above. For this reason, this channel is an independent and identically distributed (IID) channel. This assumption will again be important when we go to the asymptotic regime of a large number of uses of the channel.

Suppose that Alice and Bob just use the channel as is—Alice just sends plain bits to Bob. This scheme works reliably only if the probability of bit flip error vanishes. So, Alice and Bob could invest their best efforts into engineering the physical channel to make it reliable. But, generally, it is not possible to engineer a classical channel this way for physical or logistical reasons. For example, Alice and Bob may only have local computers at their ends and may not have access to the physical channel because the telephone company may control the channel.

Alice and Bob can employ a "systems engineering" solution to this problem rather than an engineering of the physical channel. They can redundantly encode information in a way such that Bob can have a higher probability of determining what Alice is sending, effectively reducing the level of noise on the channel. A simple example of this systems engineering solution is the three-bit majority vote code. Alice and Bob employ the following encoding:

$$0 \rightarrow 000, \qquad 1 \rightarrow 111, \tag{2.42}$$

where both '000' and '111' are *codewords*. Alice transmits the codeword '000' with three independent uses of the noisy channel if she really wants to communicate a '0' to Bob

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Channel Output	Probability
000	$(1-p)^3$
001, 010, 100	$p(1-p)^2$
011, 110, 101	$p^2(1-p)$
111	p^3

Table 2.1: The first column gives the eight possible outputs of the noisy bit-flip channel when Alice encodes a '0' with the majority vote code. The second column gives the corresponding probability of Bob receiving the particular outputs.

and she transmits the codeword '111' if she wants to send a '1' to him. The *physical* or *channel* bits are the actual bits that she transmits over the noisy channel, and the *logical* or *information* bits are those that she intends for Bob to receive. In our example, '0' is a logical bit and '000' corresponds to the physical bits.

The rate of this scheme is 1/3 because it encodes one information bit. The term "rate" is perhaps a misnomer for coding scenarios that do not involve sending bits in a time sequence over a channel. We may just as well use the majority vote code to store one bit in a memory device that may be unreliable. Perhaps a more universal term is *efficiency*. Nevertheless, we follow convention and use the term *rate* throughout this book.

Of course, the noisy bit-flip channel does not always transmit these codewords without error. So how does Bob decode in the case of error? He simply takes a *majority vote* to determine the transmitted message—he decodes as '0' if the number of zeros in the codeword he receives is greater than the number of ones.

We now analyze the performance of this simple "systems engineering" solution. Table 2.1 enumerates the probability of receiving every possible sequence of three bits, assuming that Alice transmits a '0' by encoding it as '000'. The probability of no error is $(1-p)^3$, the probability of a single-bit error is $3p(1-p)^2$, the probability of a double-bit error is $3p^2(1-p)$, and the probability of a total failure is p^3. The majority vote solution can "correct" for no error and it corrects for all single-bit errors, but it has no ability to correct for double-bit and triple-bit errors. In fact, it actually incorrectly decodes these latter two scenarios by "correcting" '011', '110', or '101' to '111' and decoding '111' as a '1'. Thus, these latter two outcomes are errors because the code has no ability to correct them. We can employ similar arguments as above to the case where Alice transmits a '1' to Bob with the majority vote code. When does this majority vote scheme perform better than no coding at all? It is exactly when the probability of error with the majority vote code is less than p, the probability of error with no coding. The probability of error is equal to the following quantity:

$$p(e) = p(e|0)p(0) + p(e|1)p(1). \tag{2.43}$$

Our analysis above suggests that the conditional probabilities $p(e|0)$ and $p(e|1)$ are equal for the majority vote code because of the symmetry in the noisy bit-flip channel. This result

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

implies that the probability of error is

$$p(e) = 3p^2(1-p) + p^3 \tag{2.44}$$
$$= 3p^2 - 2p^3, \tag{2.45}$$

because $p(0) + p(1) = 1$. We consider the following inequality to determine if the majority vote code reduces the probability of error:

$$3p^2 - 2p^3 < p. \tag{2.46}$$

This inequality simplifies as

$$0 < 2p^3 - 3p^2 + p \tag{2.47}$$
$$\therefore 0 < p(2p-1)(p-1). \tag{2.48}$$

The only values of p that satisfy the above inequality are $0 < p < 1/2$. Thus, the majority vote code reduces the probability of error only when $0 < p < 1/2$, i.e., when the noise on the channel is not too much. Too much noise has the effect of causing the codewords to flip too often, throwing off Bob's decoder.

The majority vote code gives a way for Alice and Bob to reduce the probability of error during their communication, but unfortunately, there is still a non-zero probability for the noisy channel to disrupt their communication. Is there any way that they can achieve reliable communication by reducing the probability of error to zero?

One simple approach to achieve this goal is to exploit the majority vote idea a second time. They can *concatenate* two instances of the majority vote code to produce a code with a larger number of physical bits. Concatenation consists of using one code as an "inner" code and another as an "outer" code. There is no real need for us to distinguish between the inner and outer code in this case because we use the same code for both the inner and outer code. The concatenation scheme for our case first encodes the message i, where $i \in \{0, 1\}$, using the majority vote code. Let us label the codewords as follows:

$$\bar{0} \equiv 000, \qquad \bar{1} \equiv 111. \tag{2.49}$$

For the second layer of the concatenation, we encode $\bar{0}$ and $\bar{1}$ with the majority vote code again:

$$\bar{0} \rightarrow \bar{0}\bar{0}\bar{0}, \qquad \bar{1} \rightarrow \bar{1}\bar{1}\bar{1}. \tag{2.50}$$

Thus, the overall encoding of the concatenated scheme is as follows:

$$0 \rightarrow 000\ 000\ 000, \qquad 1 \rightarrow 111\ 111\ 111. \tag{2.51}$$

The rate of the concatenated code is $1/9$ and smaller than the original rate of $1/3$. A simple application of the above performance analysis for the majority vote code shows that this concatenation scheme reduces the probability of error as follows:

$$3p^2(e) - 2p^3(e) = O\big(p^4\big). \tag{2.52}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The error probability $p(e)$ is in (2.45) and $O(p^4)$ indicates that the leading order term of the left-hand side is the fourth power in p.

The concatenated scheme achieves a lower probability of error at the cost of using more physical bits in the code. Recall that our goal is to achieve reliable communication, where there is no probability of error. A first guess for achieving reliable communication is to continue concatenating. If we concatenate again, the probability of error reduces to $O(p^6)$, and the rate drops to $1/27$. We can continue indefinitely with concatenating to make the probability of error arbitrarily small and achieve reliable communication, but the problem is that the rate approaches zero as the probability of error becomes arbitrarily small.

The above example seems to show that there is a trade-off between the rate of the encoding scheme and the desired order of error probability. Is there a way that we can code information for a noisy channel while maintaining a good rate of communication?

2.2.2 Shannon's Channel Coding Theorem

Shannon's second breakthrough coding theorem provides an affirmative answer to the above question. This answer came as a complete shock to communication researchers in 1948. Furthermore, the techniques that Shannon used in demonstrating this fact were rarely used by engineers at the time. We give a broad overview of Shannon's main idea and techniques that he used to prove his second important theorem—the noisy channel coding theorem.

2.2.3 General Model for a Channel Code

We first generalize some of the ideas in the above example. We still have Alice trying to communicate with Bob, but this time, she wants to be able to transmit a larger set of messages with asymptotically perfect reliability, rather than merely sending '0' or '1'. Suppose that she selects messages from a message set $[M]$ that consists of M messages:

$$[M] \equiv \{1, \ldots, M\}. \tag{2.53}$$

Suppose furthermore that Alice chooses a particular message m with uniform probability from the set $[M]$. This assumption of a uniform distribution for Alice's messages indicates that we do not really care much about the content of the actual message that she is transmitting. We just assume total ignorance of her message because we only really care about her ability to send any message reliably. The message set $[M]$ requires $\log(M)$ bits to represent it, where the logarithm is again base two. This number becomes important when we calculate the rate of a channel code.

The next aspect of the model that we need to generalize is the noisy channel that connects Alice to Bob. We used the bit-flip channel before, but this channel is not general enough for our purposes. A simple way to extend the channel model is to represent it as a conditional probability distribution involving an input random variable X and an output random variable Y:

$$\mathcal{N}: \qquad p_{Y|X}(y|x). \tag{2.54}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

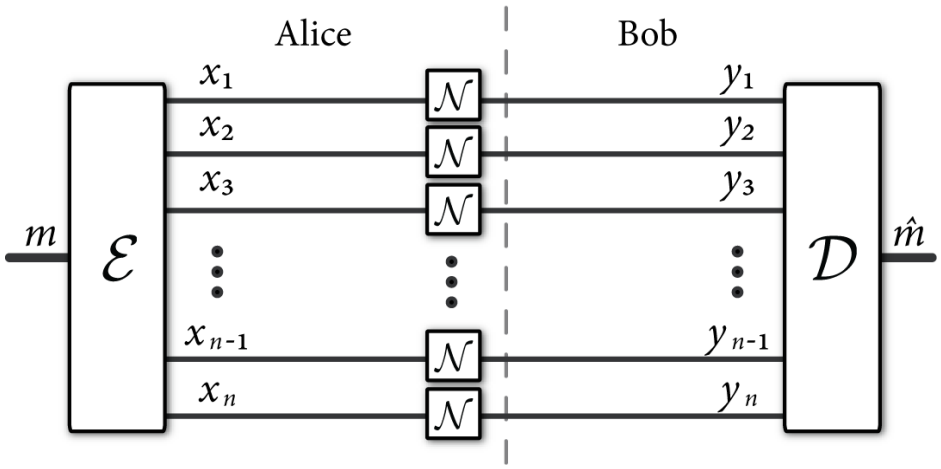

Figure 2.4: The above figure depicts Shannon's idea for a classical channel code. Alice chooses a message m with uniform probability from a message set $[M] \equiv \{1, \ldots, M\}$. She encodes the message m with an encoding operation $\mathcal{E}$. This encoding operation assigns a codeword x^n to the message m and inputs the codeword x^n to a large number of IID uses of a noisy channel $\mathcal{N}$. The noisy channel randomly corrupts the codeword x^n to a sequence y^n. Bob receives the corrupted sequence y^n and performs a decoding operation $\mathcal{D}$ to estimate the codeword x^n. This estimate of the codeword x^n then produces an estimate $\hat{m}$ of the message that Alice transmitted. A reliable code has the property that Bob can decode each message $m \in [M]$ with a vanishing probability of error when the block length n becomes large.

We use the symbol $\mathcal{N}$ to represent this more general channel model. One assumption that we make about random variables X and Y is that they are discrete, but the respective sizes of their outcome sets do not have to match. The other assumption that we make concerning the noisy channel is that it is IID. Let $X^n \equiv X_1 X_2 \cdots X_n$ and $Y^n \equiv Y_1 Y_2 \cdots Y_n$ be the random variables associated with respective sequences $x^n \equiv x_1 x_2 \cdots x_n$ and $y^n \equiv y_1 y_2 \cdots y_n$. If Alice inputs the sequence x^n to the n inputs of n respective uses of the noisy channel, a possible output sequence may be y^n. The IID assumption allows us to factor the conditional probability of the output sequence y^n:

$$p_{Y^n|X^n}(y^n|x^n) = p_{Y_1|X_1}(y_1|x_1) p_{Y_2|X_2}(y_2|x_2) \cdots p_{Y_n|X_n}(y_n|x_n) \tag{2.55}$$

$$= p_{Y|X}(y_1|x_1) p_{Y|X}(y_2|x_2) \cdots p_{Y|X}(y_n|x_n) \tag{2.56}$$

$$= \prod_{i=1}^{n} p_{Y|X}(y_i|x_i). \tag{2.57}$$

The technical name of this more general channel model is a *discrete memoryless channel*.

A coding scheme or *code* translates all of Alice's messages into codewords that can be input to n IID uses of the noisy channel. For example, suppose that Alice selects a message m to encode. We can write the codeword corresponding to message m as $x^n(m)$ because the input to the channel is some codeword that depends on m.

The last part of the model involves Bob receiving the corrupted codeword y^n over the channel and determining a potential codeword x^n with which it should be associated. We

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

do not get into any details just yet for this last decoding part—imagine for now that it operates similarly to the majority vote code example. Figure 2.4 displays Shannon's model of communication that we have described.

We calculate the *rate* of a given coding scheme as follows:

$$\text{rate} \equiv \frac{\#\text{ of message bits}}{\#\text{ of channel uses}}. \tag{2.58}$$

In our model, the rate of a given coding scheme is

$$R = \frac{1}{n}\log(M), \tag{2.59}$$

where $\log(M)$ is the number of bits needed to represent any message in the message set $[M]$ and n is the number of channel uses. The *capacity* of a noisy channel is the highest rate at which it can communicate information reliably.

We also need a way to determine the performance of any given code. Here, we list several measures of performance. Let $\mathcal{C} \equiv \{x^n(m)\}_{m\in[M]}$ represent a code that Alice and Bob choose, where $x^n(m)$ denotes each codeword corresponding to the message m. Let $p_e(m,\mathcal{C})$ denote the probability of error when Alice transmits a message $m \in [M]$ using the code $\mathcal{C}$. We denote the average probability of error as

$$\bar{p}_e(\mathcal{C}) \equiv \frac{1}{M}\sum_{m=1}^{M} p_e(m,\mathcal{C}). \tag{2.60}$$

The maximal probability of error is

$$p_e^*(\mathcal{C}) \equiv \max_{m\in[M]} p_e(m,\mathcal{C}). \tag{2.61}$$

Our ultimate aim is to make the maximal probability of error $p_e^*(\mathcal{C})$ arbitrarily small, but the average probability of error $\bar{p}_e(\mathcal{C})$ is important in the analysis. These two performance measures are related—the average probability of error is of course small if the maximal probability of error is. Perhaps surprisingly, the maximal probability is small for at least half of the messages if the average probability of error is. We make this statement more quantitative in the following exercise.

Exercise 2.2.1 Use Markov's inequality to prove that the following upper bound on the average probability of error

$$\frac{1}{M}\sum_{m} p_e(m,\mathcal{C}) \leq \epsilon \tag{2.62}$$

implies the following upper bound for at least half of the messages m:

$$p_e(m,\mathcal{C}) \leq 2\epsilon. \tag{2.63}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

You may have wondered why we use the random sequence X^n to model the inputs to the channel. We have already stated that Alice's message is a uniform random variable, and the codewords in any coding scheme directly depend on the message to be sent. For example, in the majority vote code, the channel inputs are always '000' whenever the intended message is '0' and similarly for the channel inputs '111' and the message '1'. So why is there a need to overcomplicate things by modeling the channel inputs as the random variable X^n when it seems like each codeword is a deterministic function of the intended message? We are not yet ready to answer this question but will return to it shortly.

We should also stress an important point before proceeding with Shannon's ingenious scheme for proving the existence of reliable codes for a noisy channel. In the above model, we described essentially two "layers of randomness":

1. The first layer of randomness is the uniform random variable associated with Alice's choice of a message.

2. The second layer of randomness is the noisy channel. The output of the channel is a random variable because we cannot always predict the output of the channel with certainty.

It is not possible to "play around" with these two layers of randomness. The random variable associated with Alice's message is fixed as a uniform random variable because we assume ignorance of Alice's message. The conditional probability distribution of the noisy channel is also fixed. We are assuming that Alice and Bob can learn the conditional probability distribution associated with the noisy channel by estimating it. Alternatively, we may assume that a third party has knowledge of the conditional probability distribution and informs Alice and Bob of it in some way. Regardless of how they obtain the knowledge of the distribution, we assume that they both know it and that it is fixed.

2.2.4 Description of the Proof of Shannon's Channel Coding Theorem

We are now ready to present an overview of Shannon's technique for proving the existence of a code that can achieve the capacity of a given noisy channel. Some of the methods that Shannon uses in his outline of a proof are similar to those in the first coding theorem. We again use the channel a large number of times so that the Law of Large Numbers from probability theory comes into play and allow for a small probability of error that vanishes as the number of channel uses becomes large. If the notion of typical sequences is so important in the first coding theorem, we might suspect that it should be important in the noisy channel coding theorem as well. The typical set captures a certain notion of efficiency because it is a small set when compared to the set of all sequences, but it is the set that has almost all of the probability. Thus, we should expect this efficiency to come into play somehow in the channel coding theorem.

The aspect of Shannon's technique for proving the noisy channel coding theorem that is different from the other ideas in the first theorem is the idea of *random coding*. Shannon's

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

technique adds a *third* layer of randomness to the model given above (recall that the first two are Alice's random message and the random nature of the noisy channel).

The third layer of randomness is to choose the codewords themselves in a random fashion according to a random variable X, where we choose each letter x_i of a given codeword x^n independently according to the distribution $p_X(x_i)$. It is for this reason that we model the channel inputs as a random variable. We can then write each codeword as a random variable $X^n(m)$. The probability distribution for choosing a particular codeword $x^n(m)$ is

$$\begin{aligned}\Pr\{X^n(m) = x^n(m)\} &= p_{X_1,X_2,\ldots,X_n}(x_1(m), x_2(m), \ldots, x_n(m)) && (2.64)\\ &= p_X(x_1(m))p_X(x_2(m)) \cdots p_X(x_n(m)) && (2.65)\\ &= \prod_{i=1}^{n} p_X(x_i(m)). && (2.66)\end{aligned}$$

The important result to notice is that the probability for a given codeword factors because we choose the code in an IID fashion, and perhaps more importantly, the distribution of each codeword has no explicit dependence on the message m with which it is associated. That is, the probability distribution of the first codeword is exactly the same as the probability distribution of all of the other codewords. The code $\mathcal{C}$ itself becomes a random variable in this scheme for choosing a code randomly. We now let $\mathcal{C}$ refer to the random variable that represents a random code, and we let $\mathcal{C}_0$ represent any particular deterministic code. The probability of choosing a particular code $\mathcal{C}_0 = \{x^n(m)\}_{m\in[M]}$ is

$$p_{\mathcal{C}}(\mathcal{C}_0) = \prod_{m=1}^{M} \prod_{i=1}^{n} p_X(x_i(m)), \tag{2.67}$$

and this probability distribution again has no explicit dependence on each message m in the code $\mathcal{C}_0$.

Choosing the codewords in a random way allows for a dramatic simplification in the mathematical analysis of the probability of error. Shannon's breakthrough idea was to analyze the *expectation* of the average probability of error, where the expectation is with respect to the random code $\mathcal{C}$, rather than analyzing the average probability of error itself. The expectation of the average probability of error is

$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\}. \tag{2.68}$$

This expectation is much simpler to analyze because of the random way that we choose the code. Consider that

$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\} = \mathbb{E}_{\mathcal{C}}\left\{\frac{1}{M}\sum_{m=1}^{M} p_e(m, \mathcal{C})\right\}. \tag{2.69}$$

Using linearity of the expectation, we can exchange the expectation with the sum so that

$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\mathcal{C}}\{p_e(m, \mathcal{C})\}. \tag{2.70}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Now, the expectation of the probability of error for a particular message m does not actually depend on the message m because the distribution of each random codeword $X^n(m)$ does not explicitly depend on m. This line of reasoning leads to the dramatic simplification because $\mathbb{E}_{\mathcal{C}}\{p_e(m,\mathcal{C})\}$ is then the same for all messages. So we can then say that

$$\mathbb{E}_{\mathcal{C}}\{p_e(m,\mathcal{C})\} = \mathbb{E}_{\mathcal{C}}\{p_e(1,\mathcal{C})\}. \tag{2.71}$$

(We could have equivalently chosen any message instead of the first.) We then have that

$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\} = \frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\mathcal{C}}\{p_e(1,\mathcal{C})\} \tag{2.72}$$

$$= \mathbb{E}_{\mathcal{C}}\{p_e(1,\mathcal{C})\}, \tag{2.73}$$

where the last step follows because the quantity $\mathbb{E}_{\mathcal{C}}\{p_e(1,\mathcal{C})\}$ has no dependence on m. We now only have to determine the expectation of the probability of error for *one message* instead of determining the expectation of the average error probability of the whole set. This simplification follows because random coding results in the equivalence of these two quantities.

Shannon then determined a way to obtain a bound on the the expectation of the average probability of error (we soon discuss this technique briefly) so that

$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\} \leq \epsilon, \tag{2.74}$$

where ϵ is some number that we can make arbitrarily small by letting the block size n become arbitrarily large. If it is possible to obtain a bound on the expectation of the average probability of error, then surely there exists some deterministic code $\mathcal{C}_0$ whose average probability of error meets this same bound:

$$\bar{p}_e(\mathcal{C}_0) \leq \epsilon. \tag{2.75}$$

If it were not so, then the original bound on the expectation would not be possible. This step is the *derandomization* step of Shannon's proof. Ultimately, we require a deterministic code with a high rate and arbitrarily small probability of error and this step shows the existence of such a code. The random coding technique is only useful for simplifying the mathematics of the proof.

The last step of the proof is the *expurgation* step. It is an application of the result of Exercise 2.2.1. Recall that our goal is to show the existence of a high rate code that has low maximal probability of error. But so far we only have a bound on the average probability of error. In the expurgation step, we simply throw out the half of the codewords with the worst probability of error. Throwing out the worse half of the codewords reduces the number of messages by a factor of two, but only has a negligible impact on the rate of the code. Consider that the number of messages is 2^{nR} where R is the rate of the code. Thus, the number of messages is $2^{n\left(R-\frac{1}{n}\right)}$ after throwing out the worse half of the codewords, and the rate $R-\frac{1}{n}$ is asymptotically equivalent to the rate R. After throwing out the worse half of

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

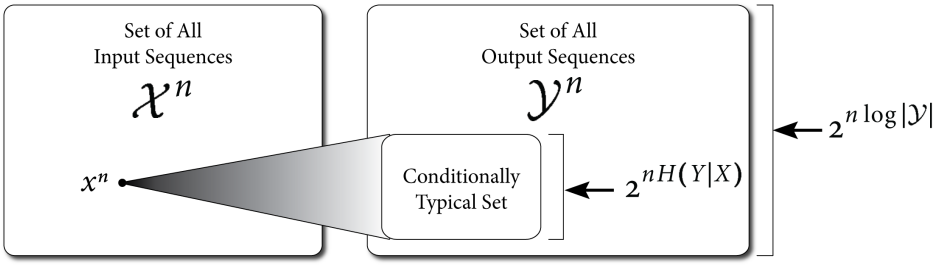

Figure 2.5: The above figure depicts the notion of a conditionally typical set. Associated to every input sequence x^n is a conditionally typical set consisting of the likely output sequences. The size of this conditionally typical set is $\approx 2^{nH(Y|X)}$. It is exponentially smaller than the set of all output sequences whenever the conditional random variable is not uniform.

the codewords, the result of Exercise 2.2.1 shows that the following bound then applies to the maximal probability of error:

$$p_e^*(\mathcal{C}_0) \leq 2\epsilon. \tag{2.76}$$

This last expurgation step ends the analysis of the probability of error.

We now discuss the size of the code that Alice and Bob employ. Recall that the rate of the code is $R = \log(M)/n$. It is convenient to define the size M of the message set $[M]$ in terms of the rate R. When we do so, the size of the message set is

$$M = 2^{nR}. \tag{2.77}$$

What is peculiar about the message set size when defined this way is that it grows exponentially with the number of channel uses. But recall that any given code exploits n channel uses to send M messages. So when we take the limit as the number of channel uses tends to infinity, we are implying that there exists a sequence of codes whose messages set size is $M = 2^{nR}$ and number of channel uses is n. We are focused on keeping the rate of the code constant and use the limit of n to make the probability of error vanish for a certain fixed rate R.

What is the maximal rate at which Alice can communicate to Bob reliably? We need to determine the number of distinguishable messages that Alice can reliably send to Bob, and we require the notion of *conditional typicality* to do so. Consider that Alice chooses codewords randomly according to random variable X with probability distribution $p_X(x)$. By the asymptotic equipartition theorem, it is highly likely that each of the codewords that Alice chooses is a typical sequence with sample entropy close to $H(X)$. In the coding scheme, Alice transmits a particular codeword x^n over the noisy channel and Bob receives a random sequence Y^n. The random sequence Y^n is a random variable that depends on x^n through the conditional probability distribution $p_{Y|X}(y|x)$. We would like a way to determine the number of possible output sequences that are likely to correspond to a particular input sequence x^n. A useful entropic quantity for this situation is the conditional entropy $H(Y|X)$, the technical details of which we leave for Chapter 10. For now, just think of this conditional entropy as

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

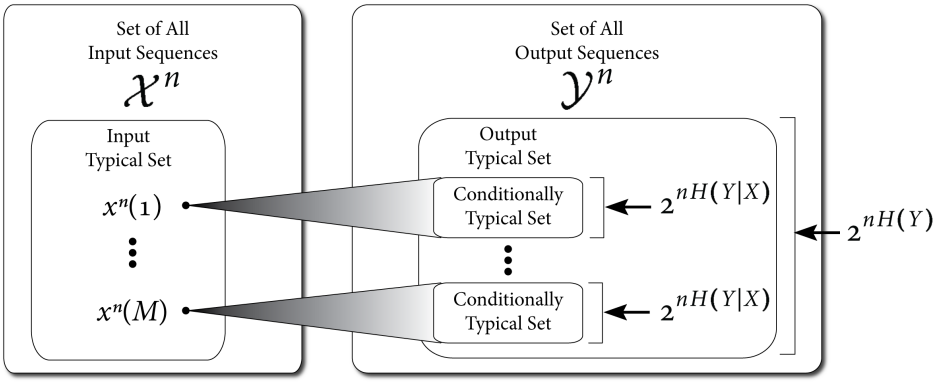

Figure 2.6: The above figure depicts the packing argument that Shannon used. The channel induces a conditionally typical set corresponding to each codeword $x^n(i)$ where $i \in \{1, \ldots, M\}$. The size of each conditionally typical output set is $2^{nH(Y|X)}$. The size of the typical set of all output sequences is $2^{nH(Y)}$. These sizes suggest that we can divide the output typical set into M conditionally typical sets and be able to distinguish $M \approx 2^{nH(Y)}/2^{nH(Y|X)}$ messages without error.

measuring the uncertainty of a random variable Y when one already knows the value of the random variable X. The conditional entropy $H(Y|X)$ is always less than the entropy $H(Y)$ unless X and Y are independent. This inequality holds because knowledge of a correlated random variable X does not increase the uncertainty about Y. It turns out that there is a notion of conditional typicality, similar to the notion of typicality, and a similar asymptotic equipartition theorem holds for conditionally typical sequences (more details in Section 13.9). This theorem also has three important properties. For each input sequence x^n, there is a corresponding conditionally typical set with the following properties:

1. It has almost all of the probability—it is highly likely that a random channel output sequence is conditionally typical given a particular input sequence.

2. Its size is $\approx 2^{nH(Y|X)}$.

3. The probability of each conditionally typical sequence y^n, given knowledge of the input sequence x^n, is $\approx 2^{-nH(Y|X)}$.

If we disregard knowledge of the input sequence used to generate an output sequence, the probability distribution that generates the output sequences is

$$p_Y(y) = \sum_x p_{Y|X}(y|x)p_X(x). \tag{2.78}$$

We can think that this probability distribution is the one that generates all the possible output sequences. The likely output sequences are in an output typical set of size $2^{nH(Y)}$.

We are now in a position to describe the structure of a random code and the size of the message set. Alice generates 2^{nR} codewords according to the distribution $p_X(x)$ and suppose

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

for now that Bob has knowledge of the code after Alice generates it. Suppose Alice sends one of the codewords over the channel. Bob is ignorant of the transmitted codeword, so from his point of view, the output sequences are generated according to the distribution $p_Y(y)$. Bob then employs typical sequence decoding. He first determines if the output sequence y^n is in the typical output set of size $2^{nH(Y)}$. If not, he declares an error. The probability of this type of error is small by the asymptotic equipartition theorem. If the output sequence y^n is in the output typical set, he uses his knowledge of the code to determine the most likely conditionally typical set of size $2^{nH(Y|X)}$ to which the output sequence belongs. If he decodes an output sequence y^n to the wrong conditionally typical set, then an error occurs. This last type of error suggests how they might structure the code in order to prevent this type of error from happening. If they structure the code so that the output conditionally typical sets do not overlap too much, then Bob should be able to decode each output sequence y^n to a unique input sequence x^n with high probability. This line of reasoning suggests that they should divide the set of output typical sequences into M sets of conditionally typical output sets, each of size $2^{nH(Y|X)}$. Thus, if they set the number of messages $M = 2^{nR}$ as follows

$$2^{nR} \approx \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y)-H(Y|X))}, \tag{2.79}$$

then our intuition is that Bob should be able to decode correctly with high probability. Such an argument is a "packing" argument because it shows how to pack information efficiently into the space of all output sequences. Figure 2.6 gives a visual depiction of the packing argument. It turns out that this intuition is correct—Alice can reliably send information to Bob if the quantity $H(Y) - H(Y|X)$ bounds the rate R:

$$R < H(Y) - H(Y|X). \tag{2.80}$$

A rate less than $H(Y) - H(Y|X)$ ensures that we can make the expectation of the average probability of error as small as we would like. We then employ the derandomization and expurgation steps, discussed before, in order to show that there exists a code whose maximal probability of error vanishes as the number n of channel uses tends to infinity.

The entropic quantity $H(Y) - H(Y|X)$ deserves special attention because it is another important entropic quantity in information theory. It is the *mutual information* between random variables X and Y and we denote it as

$$I(X;Y) \equiv H(Y) - H(Y|X). \tag{2.81}$$

It is important because it arises as the limiting rate of reliable communication. We will discuss its properties in more detail throughout this book.

There is one final step that we can take to strengthen the above coding scheme. We mentioned before that there are three layers of randomness in the coding construction: Alice's uniform choice of a message, the noisy channel, and Shannon's random coding scheme. The first two layers of randomness we do not have control over. But we actually do have control over the last layer of randomness. Alice chooses the code according to the distribution $p_X(x)$. She can choose the code according to any distribution that she would like. If she chooses it

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

according to $p_X(x)$, the resulting rate of the code is the mutual information $I(X;Y)$. We will prove later on that the mutual information $I(X;Y)$ is a concave function of the distribution $p_X(x)$ when the conditional distribution $p_{Y|X}(y|x)$ is fixed. Concavity implies that there is a unique distribution $p_X^*(x)$ that maximizes the mutual information. Thus, Alice should choose the optimum distribution $p_X^*(x)$ when she randomly generates the code, and this choice gives the largest possible rate of communication that they could have. This largest possible rate is the *capacity* of the channel and we denote it as

$$C(\mathcal{N}) \equiv \max_{p_X(x)} I(X;Y). \tag{2.82}$$

Our discussion here is just an overview of Shannon's channel capacity theorem. In Section 13.10, we give a full proof of this theorem after having developed some technical tools needed for a formal proof.

We clarify one more point. In the discussion of the operation of the code, we mentioned that Alice and Bob both have knowledge of the code. Well, how can Bob know the code if a noisy channel connects Alice to Bob? The solution to this problem is to assume that Alice and Bob have unbounded computation on their local ends. Thus, for a given code that uses the channel n times, they can both compute the above optimization problem and generate "test" codes randomly until they determine the best possible code to employ for n channel uses. They then both end up with the unique, best possible code for n uses of the given channel. This scheme might be impractical, but nevertheless, it provides a justification for both of them to have knowledge of the code that they use.

We have said before that the capacity $C(\mathcal{N})$ is the maximal rate at which Alice and Bob can communicate. But in our discussion above, we did not prove optimality—we only proved a direct coding theorem for the channel capacity theorem. It took quite some time and effort to develop this elaborate coding procedure—along the way, we repeatedly invoked one of the "elephant guns" from probability theory, the law of large numbers. It perhaps seems intuitive that typical sequence coding and decoding should lead to optimal code constructions. Typical sequences exhibit some kind of asymptotic efficiency by being the most likely to occur, but in the general case, their cardinality is exponentially smaller than the set of all sequences. But is this intuition about typical sequence coding correct? Is it possible that some other scheme for coding might beat this elaborate scheme that Shannon devised? *Without a converse theorem that proves optimality, we would never know!* If you recall from our previous discussion in Section 2.1.3 about coding theorems, we stressed how important it is to prove a converse theorem that matches the rate that the direct coding theorem suggests is optimal. For now, we delay the proof of the converse theorem because the tools for proving it are much different from the tools we described in this section. For now, accept that the formula in (2.82) is indeed the optimal rate at which two parties can communicate and we will prove this result in a later chapter.

We end the description of Shannon's channel coding theorem by giving the formal statements of the direct coding theorem and the converse theorem. The formal statement of the direct coding theorem is as follows:

If the rate of communication is less than the channel capacity, then it is possible

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

for Alice to communicate reliably to Bob, in the sense that a sequence of codes exists whose maximal probability of error vanishes as the number of channel uses tends to infinity.

The formal statement of the converse theorem is as follows:

If a reliable code exists, then the rate of this code is less than the channel capacity.

Another way of stating the converse proves to be useful later on:

If the rate of a coding scheme is greater than the channel capacity, then a reliable code does not exist, in the sense that the error probability of the coding scheme is bounded away from zero.

2.3 Summary

A general communication scenario involves one sender and one receiver. In the classical setting, we discussed two information processing tasks that they can perform. The first task was data compression or source coding, and we assumed that the sender and receiver share a noiseless classical bit channel that they use a large number of times. We can think of this noiseless classical bit channel as a *noiseless dynamic resource* that the two parties share. The resource is dynamic because we assume that there is some physical medium through which the physical carrier of information travels in order to get from the sender to the receiver. It was our aim to count the number of times they would have to use the noiseless resource in order to send information reliably. The result of Shannon's source coding theorem is that the entropy gives the minimum rate at which they have to use the noiseless resource. The second task we discussed was channel coding and we assumed that the sender and receiver share a noisy classical channel that they can use a large number of times. This noisy classical channel is a *noisy dynamic resource* that they share. We can think of this information processing task as a *simulation task*, where the goal is to simulate a noiseless dynamic resource by using a noisy dynamic resource in a redundant way. This redundancy is what allows Alice to communicate reliably to Bob, and reliable communication implies that they have effectively simulated a noiseless resource. We again had a resource count for this case, where we counted n as the number of times they use the noisy resource and nC is the number of noiseless bit channels they simulate (where C is the capacity of the channel). This notion of resource counting may not seem so important for the classical case, but it becomes much more important for the quantum case.

We now conclude our overview of Shannon's information theory. The main points to take home from this overview are the ideas that Shannon employed for constructing source and channel codes. We let the information source emit a large sequence of data, or similarly, we use the channel a large number of times so that we can invoke the law of large numbers from probability theory. The result is that we can show vanishing error for both schemes

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

by taking an asymptotic limit. In Chapter 13, we develop the theory of typical sequences in detail, proving many of the results taken for granted in this overview.

In hindsight, Shannon's methods for proving the two coding theorems are merely a *tour de force* for one idea from probability theory: the law of large numbers. Perhaps, this viewpoint undermines the contribution of Shannon, until we recall that no one had even come close to devising these methods for data compression and channel coding. The theoretical development of Shannon is one of the most important contributions to modern science because his theorems determine the ultimate rate at which we can compress and communicate information.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Part II

The Quantum Theory

CHAPTER 3

The Noiseless Quantum Theory

The simplest quantum system is the physical quantum bit or *qubit*. The qubit is a two-level quantum system—example qubit systems are the spin of an electron, the polarization of a photon, or a two-level atom with a ground state and an excited state. We do not worry too much about physical implementations in this chapter, but instead focus on the mathematical postulates of the quantum theory and operations that we can perform on qubits.

We progress from qubits to a study of physical *qudits*. Qudits are quantum systems that have d levels and are an important generalization of qubits. Again, we do not discuss physical realizations of qudits.

Noise can affect quantum systems, and we must understand methods of modeling noise in the quantum theory because our ultimate aim is to construct schemes for protecting quantum systems against the detrimental effects of noise. In Chapter 1, we remarked on the different types of noise that occur in nature. The first, and perhaps more easily comprehensible type of noise, is that which is due to our lack of information about a given scenario. We observe this type of noise in a casino, with every shuffle of cards or toss of dice. These events are random, and the random variables of probability theory model them because the outcomes are unpredictable. This noise is the same as that in all classical information processing systems. We can engineer physical systems to improve their robustness to noise.

On the other hand, the quantum theory features a fundamentally different type of noise. Quantum noise is inherent in nature and is not due to our lack of information, but is due rather to nature itself. An example of this type of noise is the "Heisenberg noise" that results from the uncertainty principle. If we know the momentum of a given particle from performing a precise measurement of it, then we know absolutely nothing about its position—a measurement of its position gives a random result. Similarly, if we know the rectilinear polarization of a photon by precisely measuring it, then a future measurement of its diagonal polarization will give a random result. It is important to keep the distinction clear between these two types of noise.

We explore the postulates of the quantum theory in this chapter, by paying particular attention to qubits. These postulates apply to a closed quantum system that is isolated from everything else in the universe. We label this first chapter "Noiseless Quantum Theory"

because closed quantum systems do not interact with their surroundings and are thus not subject to corruption and information loss. Interaction with surrounding systems can lead to loss of information in the sense of the classical noise that we described above. Closed quantum systems do undergo a certain type of quantum noise, such as that from the uncertainty principle and the act of measurement, because they are subject to the postulates of the quantum theory. The name "Noiseless Quantum Theory" thus indicates the closed, ideal nature of the quantum systems discussed in this chapter.

This chapter introduces the four postulates of the quantum theory. The mathematical tools of the quantum theory rely on the fundamentals of linear algebra—vectors and matrices of complex numbers. It may seem strange at first that we need to incorporate the machinery of linear algebra in order to describe a physical system in the quantum theory, but it turns out that this description uses the simplest set of mathematical tools to predict the phenomena that a quantum system exhibits. The hallmark of the quantum theory is that certain operations do not commute with one another, and matrices are the simplest mathematical objects that capture this idea of noncommutativity.

3.1 Overview

We first briefly overview how information is processed with quantum systems. This usually consists of three steps: state preparation, quantum operations, and measurement. State preparation is where we initialize a quantum system to some beginning state, depending on what operation we would like a quantum system to execute. There could be some classical control device that initializes the state of the quantum system. Observe that the input system for this step is a classical system, and the output system is quantum. After initializing the state of the quantum system, we perform some quantum operations that evolve its state. This stage is where we can take advantage of quantum effects for enhanced information processing abilities. Both the input and output systems of this step are quantum. Finally, we need some way of reading out the result of the computation, and we can do so with a measurement. The input system for this step is quantum, and the output is classical. Figure 3.1 depicts all of these steps. In a quantum communication protocol, spatially separated parties may execute different parts of these steps, and we are interested in keeping track of the nonlocal resources needed to implement a communication protocol. Section 3.2 describes quantum states (and thus state preparation), Section 3.3 describes the noiseless evolution of quantum states, and Section 3.4 describes "read out" or measurement. For now, we assume that we can perform all of these steps perfectly and later chapters discuss how to incorporate the effects of noise.

3.2 Quantum Bits

The simplest quantum system is a two-state system: a physical qubit. Let $|0\rangle$ denote one possible state of the system. The left vertical bar and the right angle bracket indicate that we

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

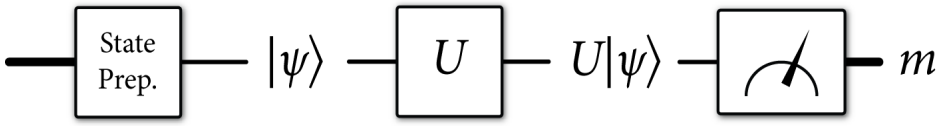

Figure 3.1: All of the steps in a typical noiseless quantum information processing protocol. A classical control (depicted by the thick black line on the left) initializes the state of a quantum system. The quantum system then evolves according to some unitary operation (described in Section 3.3). The final step is a measurement that reads out some classical data m from the quantum system.

are using the Dirac notation to represent this state. The Dirac notation has some advantages for performing calculations in the quantum theory, and we highlight some of these advantages as we progress through our development. Let $|1\rangle$ denote another possible state of the qubit. We can encode a classical bit or *cbit* into a qubit with the following mapping:

$$0 \rightarrow |0\rangle, \qquad 1 \rightarrow |1\rangle. \tag{3.1}$$

So far, nothing in our description above distinguishes a classical bit from a qubit, except for the funny vertical bar and angle bracket that we place around the bit values. The quantum theory predicts that the above states are not the only possible states of a qubit. Arbitrary *superpositions* (linear combinations) of the above two states are possible as well because the quantum theory is a linear theory. Suffice it to say that the linearity of the quantum theory results from the linearity of Schrödinger's equation that governs the evolution of quantum systems.[1] A general noiseless qubit can be in the following state:

$$\alpha|0\rangle + \beta|1\rangle, \tag{3.2}$$

where the coefficients α and β are arbitrary complex numbers with unit norm:

$$|\alpha|^2 + |\beta|^2 = 1. \tag{3.3}$$

The coefficients α and β are *probability amplitudes*—they are not probabilities themselves but allow us to calculate probabilities. The unit-norm constraint results from the *Born rule* (the probabilistic interpretation) of the quantum theory, and we speak more on this constraint and probability amplitudes when we introduce the measurement postulate.

The possibility of superposition states indicates that we cannot represent the states $|0\rangle$ and $|1\rangle$ with the Boolean algebra of the respective classical bits 0 and 1 because Boolean algebra does not allow for superposition states. We instead require the mathematics of *linear algebra* to describe these states. It is beneficial at first to define a vector representation of

[1] We will not present Schrödinger's equation in this book, but instead focus on a "quantum information" presentation of the quantum theory. Griffith's book on quantum mechanics introduces the quantum theory from the Schrödinger equation if you are interested [116].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

the states $|0\rangle$ and $|1\rangle$:

$$|0\rangle \equiv \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{3.4}$$

$$|1\rangle \equiv \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{3.5}$$

The $|0\rangle$ and $|1\rangle$ states are called "kets" in the language of the Dirac notation, and it is best at first to think of them merely as column vectors. The superposition state in (3.2) then has a representation as the following two-dimensional vector:

$$|\psi\rangle \equiv \alpha|0\rangle + \beta|1\rangle = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{3.6}$$

The representation of quantum states with vectors is helpful in understanding some of the mathematics that underpins the theory, but it turns out to be much more useful for our purposes to work directly with the Dirac notation. We give the vector representation for now, but later on, we will only employ the Dirac notation.

The *Bloch sphere*, depicted in Figure 3.2, gives a valuable way to visualize a qubit. Consider any two qubits that are equivalent up to a differing global phase. For example, these two qubits could be

$$|\psi_0\rangle \equiv |\psi\rangle, \qquad |\psi_1\rangle \equiv e^{i\chi}|\psi\rangle, \tag{3.7}$$

where $0 \leq \chi < 2\pi$. There is a sense in which these two qubits are physically equivalent because they give the same physical results when we measure them (more on this point when we introduce the measurement postulate in Section 3.4). Suppose that the probability amplitudes α and β have the following respective representations as complex numbers:

$$\alpha = r_0 e^{i\varphi_0}, \tag{3.8}$$

$$\beta = r_1 e^{i\varphi_1}. \tag{3.9}$$

We can factor out the phase $e^{i\varphi_0}$ from both coefficients α and β, and we still have a state that is physically equivalent to the state in (3.2):

$$|\psi\rangle \equiv r_0|0\rangle + r_1 e^{i(\varphi_1 - \varphi_0)}|1\rangle, \tag{3.10}$$

where we redefine $|\psi\rangle$ to represent the state because of the equivalence mentioned in (3.7). Let $\varphi \equiv \varphi_1 - \varphi_0$, where $0 \leq \varphi < 2\pi$. Recall that the unit-norm constraint requires $|r_0|^2 + |r_1|^2 = 1$. We can thus parametrize the values of r_0 and r_1 in terms of one parameter θ so that

$$r_0 = \cos(\theta/2), \tag{3.11}$$

$$r_1 = \sin(\theta/2). \tag{3.12}$$

The parameter θ varies between 0 and π. This range of θ and the factor of two give a unique representation of the qubit. One may think to have θ vary between 0 and 2π and omit the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

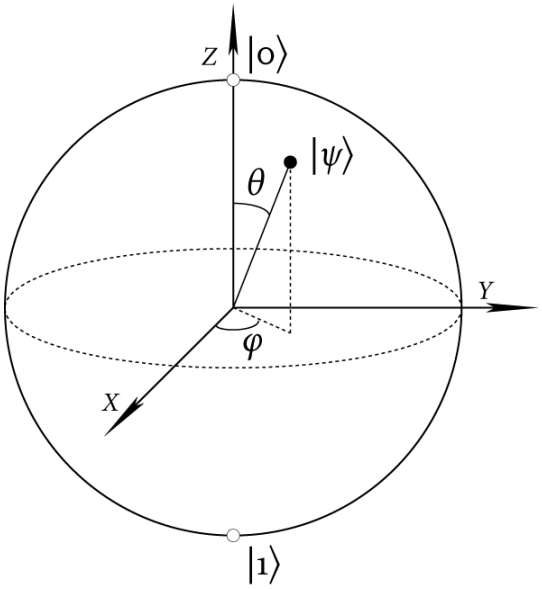

Figure 3.2: The Bloch sphere representation of a qubit. Any qubit $|\psi\rangle$ admits a representation in terms of two angles θ and φ where $0 \leq \theta \leq \pi$ and $0 \leq \varphi < 2\pi$. The state of any qubit in terms of these angles is $|\psi\rangle = \cos(\theta/2)|0\rangle + e^{i\varphi}\sin(\theta/2)|1\rangle$.

factor of two, but this parametrization would not uniquely characterize the qubit in terms of the parameters θ and φ. The parametrization in terms of θ and φ gives the Bloch sphere representation of the qubit in (3.2):

$$|\psi\rangle \equiv \cos(\theta/2)|0\rangle + \sin(\theta/2)e^{i\varphi}|1\rangle. \tag{3.13}$$

We can plot the state of any qubit on a unit sphere, called the Bloch sphere. Figure 3.2 depicts this representation of a qubit.

In linear algebra, column vectors are not the only type of vectors—row vectors are useful as well. Is there an equivalent of a row vector in Dirac notation? The Dirac notation provides an entity called a "bra," that has a representation as a row vector. The bras corresponding to the kets $|0\rangle$ and $|1\rangle$ are as follows:

$$\langle 0| \equiv \begin{bmatrix} 1 & 0 \end{bmatrix}, \tag{3.14}$$

$$\langle 1| \equiv \begin{bmatrix} 0 & 1 \end{bmatrix}, \tag{3.15}$$

and are the matrix conjugate transpose of the kets $|0\rangle$ and $|1\rangle$:

$$\langle 0| = (|0\rangle)^\dagger, \tag{3.16}$$

$$\langle 1| = (|1\rangle)^\dagger. \tag{3.17}$$

We require the conjugate transpose operation (as opposed to just the transpose) because the mathematical representation of a general quantum state can have complex entries.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The bras do not represent quantum states, but are helpful in calculating probability amplitudes. For our example qubit in (3.2), suppose that we would like to determine the probability amplitude that the state is $|0\rangle$. We can combine the state in (3.2) with the bra $\langle 0|$ as follows:

$$\begin{aligned}
\langle 0||\psi\rangle &= \langle 0|(\alpha|0\rangle + \beta|1\rangle) && (3.18)\\
&= \alpha\langle 0||0\rangle + \beta\langle 0||1\rangle && (3.19)\\
&= \alpha\begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \beta\begin{bmatrix} 1 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} && (3.20)\\
&= \alpha \cdot 1 + \beta \cdot 0 && (3.21)\\
&= \alpha. && (3.22)
\end{aligned}$$

The above calculation may seem as if it is merely an exercise in linear algebra, with a "glorified" Dirac notation, but it is a standard calculation in the quantum theory. A quantity like $\langle 0||\psi\rangle$ occurs so often in the quantum theory that we abbreviate it as

$$\langle 0|\psi\rangle \equiv \langle 0||\psi\rangle, \tag{3.23}$$

and the above notation is known as a "braket."[2] The physical interpretation of the quantity $\langle 0|\psi\rangle$ is that it is the probability amplitude for being in the state $|0\rangle$, and likewise, the quantity $\langle 1|\psi\rangle$ is the probability amplitude for being in the state $|1\rangle$. We can also determine that the amplitude $\langle 1|0\rangle$ (for the state $|0\rangle$ to be in the state $|1\rangle$) and the amplitude $\langle 0|1\rangle$ are both equal to zero. These two states are *orthogonal states* because they have no overlap. The amplitudes $\langle 0|0\rangle$ and $\langle 1|1\rangle$ are both equal to one by following a similar calculation.

Our next task may seem like a frivolous exercise, but we would like to determine the amplitude for any state $|\psi\rangle$ to be in the state $|\psi\rangle$, i.e., to be itself. Following the above method, this amplitude is $\langle\psi|\psi\rangle$ and we calculate it as

$$\begin{aligned}
\langle\psi|\psi\rangle &= (\langle 0|\alpha^* + \langle 1|\beta^*)(\alpha|0\rangle + \beta|1\rangle) && (3.24)\\
&= \alpha^*\alpha\langle 0|0\rangle + \beta^*\alpha\langle 1|0\rangle + \alpha^*\beta\langle 0|1\rangle + \beta^*\beta\langle 1|1\rangle && (3.25)\\
&= |\alpha|^2 + |\beta|^2 && (3.26)\\
&= 1, && (3.27)
\end{aligned}$$

where we have used the orthogonality relations of $\langle 0|0\rangle$, $\langle 1|0\rangle$, $\langle 0|1\rangle$, and $\langle 1|1\rangle$, and the unit-norm constraint. We come back to the unit-norm constraint in our discussion of quantum measurement, but for now, we have shown that any quantum state has a unit amplitude for being itself.

The states $|0\rangle$ and $|1\rangle$ are a particular basis for a qubit that we call the *computational basis*. The computational basis is the standard basis that we employ in quantum computation

[2]It is for this (silly) reason that Dirac decided to use the names "bra" and "ket," because putting them together gives a "braket." The names in the notation may be silly, but the notation itself has persisted over time because this way of representing quantum states turns out to be useful. We will avoid the use of the terms "bra" and "ket" as much as we can, only resorting to these terms if necessary.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and communication, but other bases are important as well. Consider that the following two vectors form an orthonormal basis:

$$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{3.28}$$

The above alternate basis is so important in quantum information theory that we define a Dirac notation shorthand for it, and we can also define the basis in terms of the computational basis:

$$|+\rangle \equiv \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \tag{3.29}$$

$$|-\rangle \equiv \frac{|0\rangle - |1\rangle}{\sqrt{2}}. \tag{3.30}$$

The common names for this alternate basis are the "$+/-$" basis, the Hadamard basis, or the diagonal basis. It is preferable for us to use the Dirac notation, but we are using the vector representation as an aid for now.

Exercise 3.2.1 Determine the Bloch sphere angles θ and φ for the states $|+\rangle$ and $|-\rangle$.

What is the amplitude that the state in (3.2) is in the state $|+\rangle$? What is the amplitude that it is in the state $|-\rangle$? These are questions to which the quantum theory provides simple answers. We employ the bra $\langle +|$ and calculate the amplitude $\langle +|\psi\rangle$ as

$$\langle +|\psi\rangle = \langle +|(\alpha|0\rangle + \beta|1\rangle) \tag{3.31}$$

$$= \alpha\langle +|0\rangle + \beta\langle +|1\rangle \tag{3.32}$$

$$= \frac{\alpha + \beta}{\sqrt{2}}. \tag{3.33}$$

The result follows by employing the definition in (3.29) and doing similar linear algebraic calculations as the example in (3.22). We can also calculate the amplitude $\langle -|\psi\rangle$ as

$$\langle -|\psi\rangle = \frac{\alpha - \beta}{\sqrt{2}}. \tag{3.34}$$

The above calculation follows from similar manipulations.

The $+/-$ basis is a *complete* orthonormal basis, meaning that we can represent any qubit state in terms of the two basis states $|+\rangle$ and $|-\rangle$. Indeed, the above probability amplitude calculations suggest that we can represent the qubit in (3.2) as the following superposition state:

$$|\psi\rangle = \left(\frac{\alpha + \beta}{\sqrt{2}}\right)|+\rangle + \left(\frac{\alpha - \beta}{\sqrt{2}}\right)|-\rangle. \tag{3.35}$$

The above representation is an alternate one if we would like to "see" the qubit state represented in the $+/-$ basis. We can substitute the equivalences in (3.33) and (3.34) to represent the state $|\psi\rangle$ as

$$|\psi\rangle = \langle +|\psi\rangle|+\rangle + \langle -|\psi\rangle|-\rangle. \tag{3.36}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The amplitudes $\langle +|\psi\rangle$ and $\langle -|\psi\rangle$ are both scalar quantities so that the above quantity is equivalent to the following one:

$$|\psi\rangle = |+\rangle\langle +|\psi\rangle + |-\rangle\langle -|\psi\rangle. \tag{3.37}$$

The order of the multiplication in the terms $|+\rangle\langle +|\psi\rangle$ and $|-\rangle\langle -|\psi\rangle$ does not matter, i.e., the following equivalence holds

$$|+\rangle(\langle +|\psi\rangle) = (|+\rangle\langle +|)|\psi\rangle, \tag{3.38}$$

and the same for $|-\rangle\langle -|\psi\rangle$. The quantity on the left is a ket multiplied by an amplitude, whereas the quantity on the right is a linear operator multiplying a ket, but linear algebra tells us that these two quantities are equivalent. The operators $|+\rangle\langle +|$ and $|-\rangle\langle -|$ are special operators—they are rank-one projection operators, meaning that they project onto a one-dimensional subspace. Using linearity, we have the following equivalence:

$$|\psi\rangle = (|+\rangle\langle +| + |-\rangle\langle -|)|\psi\rangle. \tag{3.39}$$

The above equation indicates a seemingly trivial, but important point—the operator $|+\rangle\langle +| + |-\rangle\langle -|$ is equivalent to the identity operator and we can write

$$I = |+\rangle\langle +| + |-\rangle\langle -|, \tag{3.40}$$

where I stands for the identity operator. This relation is known as the *completeness relation* or the *resolution of the identity*. Given any orthonormal basis, we can always construct a resolution of the identity by summing over the rank-one projection operators formed from each of the orthonormal basis states. For example, the computational basis states give another way to form a resolution of the identity operator:

$$I = |0\rangle\langle 0| + |1\rangle\langle 1|. \tag{3.41}$$

This simple trick provides a way to find the representation of a quantum state in any basis.

3.3 Reversible Evolution

Physical systems evolve as time progresses. The application of a magnetic field to an electron can change its spin and pulsing an atom with a laser can excite one of its electrons from a ground state to an excited state. These are only a couple of ways in which physical systems can change.

The Schrödinger equation governs the evolution of a closed quantum system. In this book, we will not even state the Schrödinger equation, but we will instead focus on its major result. *The evolution of a closed quantum system is reversible if we do not learn anything about the state of the system.* Reversibility implies that we can determine the input state

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

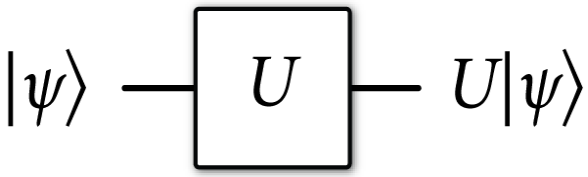

Figure 3.3: The above figure is a quantum circuit diagram that depicts the evolution of a quantum state $|\psi\rangle$ according to a unitary operator U.

of an evolution given the output state and knowledge of the evolution. An example of a single-qubit reversible operation is a NOT gate:

$$|0\rangle \rightarrow |1\rangle, \qquad |1\rangle \rightarrow |0\rangle. \tag{3.42}$$

In the classical world, we would say that the NOT gate merely flips the value of the input classical bit. In the quantum world, the NOT gate flips the basis states $|0\rangle$ and $|1\rangle$. The NOT gate is reversible because we can simply apply the NOT gate again to recover the original input state—the NOT gate is its own inverse.

In general, a closed quantum system evolves according to a unitary operator U. Unitary evolution implies reversibility because a unitary operator always possesses an inverse—its inverse is merely $U^\dagger$. This property gives the relations:

$$U^\dagger U = UU^\dagger = I. \tag{3.43}$$

The unitary property also ensures that evolution preserves the unit-norm constraint (an important requirement for a physical state that we discuss in the section on measurement). Consider applying the unitary operator U to the example qubit state in (3.2):

$$U|\psi\rangle. \tag{3.44}$$

Figure 3.3 depicts a quantum circuit diagram for unitary evolution.

The bra that is dual to the above state is $\langle\psi|U^\dagger$ (we again apply the conjugate transpose operation to get the bra). We showed in (3.24-3.27) that every quantum state should have a unit amplitude for being itself. This relation holds for the state $U|\psi\rangle$ because the operator U is unitary:

$$\langle\psi|U^\dagger U|\psi\rangle = \langle\psi|I|\psi\rangle = \langle\psi|\psi\rangle = 1. \tag{3.45}$$

The assumption that a vector always has a unit amplitude for being itself is one of the crucial assumptions of the quantum theory, and the above reasoning demonstrates that unitary evolution complements this assumption.

3.3.1 Matrix Representations of Operators

We now explore some properties of the NOT gate. Let X denote the operator corresponding to a NOT gate. The action of X on the computational basis states is as follows:

$$X|i\rangle = |i \oplus 1\rangle, \tag{3.46}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $i = \{0, 1\}$ and $\oplus$ denotes binary addition. Suppose the NOT gate acts on a superposition state:

$$X(\alpha|0\rangle + \beta|1\rangle) \tag{3.47}$$

By the linearity of the quantum theory, the X operator distributes so that the above expression is equal to the following one:

$$\alpha X|0\rangle + \beta X|1\rangle = \alpha|1\rangle + \beta|0\rangle. \tag{3.48}$$

Indeed, the NOT gate X merely flips the basis states of any quantum state when represented in the computational basis.

We can determine a *matrix representation* for the operator X by using the bras $\langle 0|$ and $\langle 1|$. Consider the relations in (3.46). Let us combine the relations with the bra $\langle 0|$:

$$\langle 0|X|0\rangle = \langle 0|1\rangle = 0, \tag{3.49}$$

$$\langle 0|X|1\rangle = \langle 0|0\rangle = 1. \tag{3.50}$$

Likewise, we can combine with the bra $\langle 1|$:

$$\langle 1|X|0\rangle = \langle 1|1\rangle = 1, \tag{3.51}$$

$$\langle 1|X|1\rangle = \langle 1|0\rangle = 0. \tag{3.52}$$

We can place these entries in a matrix to give a matrix representation of the operator X:

$$\begin{bmatrix} \langle 0|X|0\rangle & \langle 0|X|1\rangle \\ \langle 1|X|0\rangle & \langle 1|X|1\rangle \end{bmatrix}, \tag{3.53}$$

where we order the rows according to the bras and order the columns according to the kets. We then say that

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{3.54}$$

and adopt the convention that the symbol X refers to both the operator X and its matrix representation (this is an abuse of notation, but it should be clear from context when X refers to an operator and when it refers to the matrix representation of the operator).

Let us now observe some uniquely quantum behavior. We would like to consider the action of the NOT operator X on the $+/-$ basis. First, let us consider what happens if we operate on the $|+\rangle$ state with the X operator. Recall that the state $|+\rangle = 1/\sqrt{2}(|0\rangle + |1\rangle)$ so that

$$X|+\rangle = X\left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right) \tag{3.55}$$

$$= \frac{X|0\rangle + X|1\rangle}{\sqrt{2}} \tag{3.56}$$

$$= \frac{|1\rangle + |0\rangle}{\sqrt{2}} \tag{3.57}$$

$$= |+\rangle. \tag{3.58}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above development shows that the state $|+\rangle$ is a special state with respect to the NOT operator X—it is an *eigenstate* of X with *eigenvalue* one. An eigenstate of an operator is one that is invariant under the action of the operator. The coefficient in front of the eigenstate is the *eigenvalue* corresponding to the eigenstate. Under a unitary evolution, the coefficient in front of the eigenstate is just a complex phase, but this global phase has no effect on the observations resulting from a measurement of the state because two quantum states are equivalent up to a differing global phase.

Now, let us consider the action of the NOT operator X on the state $|-\rangle$. Recall that $|-\rangle \equiv 1/\sqrt{2}(|0\rangle - |1\rangle)$. Calculating similarly, we get that

$$X|-\rangle = X\left(\frac{|0\rangle - |1\rangle}{\sqrt{2}}\right) \tag{3.59}$$

$$= \frac{X|0\rangle - X|1\rangle}{\sqrt{2}} \tag{3.60}$$

$$= \frac{|1\rangle - |0\rangle}{\sqrt{2}} \tag{3.61}$$

$$= -|-\rangle. \tag{3.62}$$

So the state $|-\rangle$ is also an eigenstate of the operator X, but its eigenvalue is -1.

We can find a matrix representation of the X operator in the $+/-$ basis as well:

$$\begin{bmatrix} \langle +|X|+\rangle & \langle +|X|-\rangle \\ \langle -|X|+\rangle & \langle -|X|-\rangle \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{3.63}$$

This representation demonstrates that the X operator is diagonal with respect to the $+/-$ basis, and therefore, the $+/-$ basis is an *eigenbasis* for the X operator. It is always handy to know the eigenbasis of a unitary operator U because this eigenbasis gives the states that are invariant under an evolution according to U.

Let Z denote the operator that flips states in the $+/-$ basis:

$$Z|+\rangle \rightarrow |-\rangle, \qquad Z|-\rangle \rightarrow |+\rangle. \tag{3.64}$$

Using an analysis similar to that which we did for the X operator, we can find a matrix representation of the Z operator in the $+/-$ basis:

$$\begin{bmatrix} \langle +|Z|+\rangle & \langle +|Z|-\rangle \\ \langle -|Z|+\rangle & \langle -|Z|-\rangle \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{3.65}$$

Interestingly, the matrix representation for the Z operator in the $+/-$ basis is the same as that for the X operator in the computational basis. For this reason, we call the Z operator the *phase flip* operator.[3]

[3] A more appropriate name might be the "bit flip in the $+/-$ basis operator," but this name is too long, so we stick with the term "phase flip."

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We expect the following steps to hold because the quantum theory is a linear theory:

$$Z\left(\frac{|+\rangle + |-\rangle}{\sqrt{2}}\right) = \frac{Z|+\rangle + Z|-\rangle}{\sqrt{2}} = \frac{|-\rangle + |+\rangle}{\sqrt{2}} = \frac{|+\rangle + |-\rangle}{\sqrt{2}}, \tag{3.66}$$

$$Z\left(\frac{|+\rangle - |-\rangle}{\sqrt{2}}\right) = \frac{Z|+\rangle - Z|-\rangle}{\sqrt{2}} = \frac{|-\rangle - |+\rangle}{\sqrt{2}} = -\left(\frac{|+\rangle - |-\rangle}{\sqrt{2}}\right). \tag{3.67}$$

The above steps demonstrate that the states $1/\sqrt{2}(|+\rangle + |-\rangle)$ and $1/\sqrt{2}(|+\rangle - |-\rangle)$ are both eigenstates of the Z operators. These states are none other than the respective computational basis states $|0\rangle$ and $|1\rangle$, by inspecting the definitions in (3.29-3.30) of the $+/-$ basis. Thus, a matrix representation of the Z operator in the computational basis is

$$\begin{bmatrix} \langle 0|Z|0\rangle & \langle 0|Z|1\rangle \\ \langle 1|Z|0\rangle & \langle 1|Z|1\rangle \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \tag{3.68}$$

and is a diagonalization of the operator Z. So, the behavior of the Z operator in the computational basis is the same as the behavior of the X operator in the $+/-$ basis.

3.3.2 Commutators and Anticommutators

The *commutator* $[A, B]$ of two operators A and B is as follows:

$$[A, B] \equiv AB - BA. \tag{3.69}$$

Two operators commute if and only if their commutator is equal to zero.

The *anticommutator* $\{A, B\}$ of two operators A and B is as follows:

$$\{A, B\} \equiv AB + BA. \tag{3.70}$$

We say that two operators *anticommute* if their anticommutator is equal to zero.

Exercise 3.3.1 Find a matrix representation for $[X, Z]$ in the basis $\{|0\rangle, |1\rangle\}$.

3.3.3 The Pauli Matrices

The convention in quantum theory is to take the computational basis as the *standard basis* for representing physical qubits. The standard matrix representation for the above two operators is as follows when we choose the computational basis as the standard basis:

$$X \equiv \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Z \equiv \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{3.71}$$

The identity operator I has the following representation in any basis:

$$I \equiv \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3.72}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Another operator, the Y operator, is a useful one to consider as well. The Y operator has the following matrix representation in the computational basis:

$$Y \equiv \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}. \tag{3.73}$$

It is easy to check that $Y = iXZ$, and for this reason, we can think of the Y operator as a combined bit and phase flip. The four matrices I, X, Y, and Z are special for the manipulation of physical qubits and are known as the *Pauli matrices*.

Exercise 3.3.2 Show that the Pauli matrices are all Hermitian, unitary, they square to the identity, and their eigenvalues are ± 1.

Exercise 3.3.3 Represent the eigenstates of the Y operator in the computational basis.

Exercise 3.3.4 Show that the Pauli matrices either commute or anticommute.

Exercise 3.3.5 Let us label the Pauli matrices as $\sigma_0 \equiv I$, $\sigma_1 \equiv X$, $\sigma_2 \equiv Y$, and $\sigma_3 \equiv Z$, Show that $\text{Tr}\{\sigma_i \sigma_j\} = 2\delta_{ij}$ for all $i, j \in \{0, \ldots, 3\}$.

3.3.4 Hadamard Gate

Another important unitary operator is the transformation that takes the computational basis to the $+/-$ basis. This transformation is the Hadamard transformation:

$$|0\rangle \rightarrow |+\rangle, \tag{3.74}$$
$$|1\rangle \rightarrow |-\rangle. \tag{3.75}$$

Using the above relations, we can represent the Hadamard transformation as the following operator:

$$H \equiv |+\rangle\langle 0| + |-\rangle\langle 1|. \tag{3.76}$$

It is straightforward to check that the above operator implements the transformation in (3.75).

Now consider a generalization of the above construction. Suppose that one orthonormal basis is $\{|\psi_i\rangle\}_{i\in\{0,1\}}$ and another is $\{|\phi_i\rangle\}_{i\in\{0,1\}}$ where the index i merely indexes the states in each orthonormal basis. Then the unitary operator that takes states in the first basis to states in the second basis is

$$\sum_{i=0,1} |\phi_i\rangle\langle\psi_i|. \tag{3.77}$$

Exercise 3.3.6 Show that the Hadamard operator H has the following matrix representation in the computational basis:

$$H \equiv \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{3.78}$$

Exercise 3.3.7 Show that the Hadamard operator is its own inverse by employing the above matrix representation and by using its operator form in (3.76).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

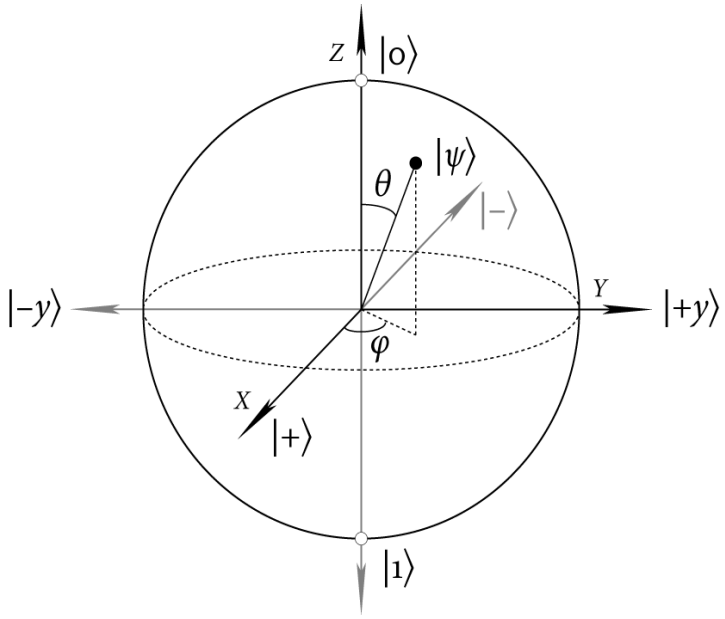

Figure 3.4: The above figure provides more labels for states on the Bloch sphere. The Z axis has its points on the sphere as eigenstates of the Pauli Z operator, the X axis has eigenstates of the Pauli X operator, and the Y axis has eigenstates of the Pauli Y operator. The rotation operators $R_X(\phi)$, $R_Y(\phi)$, and $R_Z(\phi)$ rotate a state on the sphere by an angle ϕ about the respective X, Y, and Z axis.

Exercise 3.3.8 If the Hadamard gate is its own inverse, then it takes the states $|+\rangle$ and $|-\rangle$ to the respective states $|0\rangle$ and $|1\rangle$ and we can represent it as the following operator:

$$H = |0\rangle\langle +| + |1\rangle\langle -|. \tag{3.79}$$

Show that

$$|0\rangle\langle +| + |1\rangle\langle -| = |+\rangle\langle 0| + |-\rangle\langle 1|. \tag{3.80}$$

Exercise 3.3.9 Show that $HXH = Z$ and that $HZH = X$.

3.3.5 Rotation Operators

We end this section on the evolution of quantum states by discussing "rotation evolutions" and by giving a more complete picture of the Bloch sphere. The rotation operators $R_X(\phi)$, $R_Y(\phi)$, $R_Z(\phi)$ are functions of the respective Pauli operators X, Y, Z where

$$R_X(\phi) \equiv \exp\{iX\phi/2\}, \tag{3.81}$$
$$R_Y(\phi) \equiv \exp\{iY\phi/2\}, \tag{3.82}$$
$$R_Z(\phi) \equiv \exp\{iZ\phi/2\}, \tag{3.83}$$

and ϕ is some angle such that $0 \leq \phi < 2\pi$. How do we determine a function of an operator? The standard way is to represent the operator in its diagonal basis and apply the function to

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

the eigenvalues of the operator. For example, the diagonal representation of the X operator is

$$X = |+\rangle\langle+| - |-\rangle\langle-|. \tag{3.84}$$

Applying the function $\exp\{iX\phi/2\}$ to the eigenvalues of X gives

$$R_X(\phi) = \exp\{i\phi/2\}|+\rangle\langle+| + \exp\{-i\phi/2\}|-\rangle\langle-|. \tag{3.85}$$

More generally, suppose that an Hermitian operator A has the spectral decomposition $A = \sum_i a_i|i\rangle\langle i|$ for some orthonormal basis $\{|i\rangle\}$. Then the operator $f(A)$ for some function f is as follows:

$$f(A) = \sum_i f(a_i)|i\rangle\langle i|. \tag{3.86}$$

Exercise 3.3.10 Show that the rotation operators $R_X(\phi)$, $R_Y(\phi)$, $R_Z(\phi)$ are equivalent to the following expressions:

$$R_X(\phi) = \cos(\phi/2)I + i\sin(\phi/2)X, \tag{3.87}$$
$$R_Y(\phi) = \cos(\phi/2)I + i\sin(\phi/2)Y, \tag{3.88}$$
$$R_Z(\phi) = \cos(\phi/2)I + i\sin(\phi/2)Z, \tag{3.89}$$

by using the facts that

$$\cos(\phi/2) = \frac{1}{2}\big(e^{i\phi/2} + e^{-i\phi/2}\big), \tag{3.90}$$
$$\sin(\phi/2) = \frac{1}{2i}\big(e^{i\phi/2} - e^{-i\phi/2}\big). \tag{3.91}$$

Figure 3.4 provides a more detailed picture of the Bloch sphere since we have now established the Pauli operators and their eigenstates. The computational basis states are the eigenstates of the Z operator and are the north and south poles on the Bloch sphere. The $+/-$ basis states are the eigenstates of the X operator and the calculation from Exercise 3.2.1 shows that they are the "east and west poles" of the Bloch sphere. We leave it as another exercise to show that the Y eigenstates are the other poles along the equator of the Bloch sphere.

Exercise 3.3.11 Determine the Bloch sphere angles θ and φ for the eigenstates of the Pauli Y operator.

3.4 Measurement

Measurement is another type of evolution that a quantum system can undergo, and this type of evolution is unique to the quantum theory. It is an evolution that allows us to retrieve classical information from a quantum state and thus is the way that we can "read out" information. Suppose that we would like to learn something about the quantum state

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

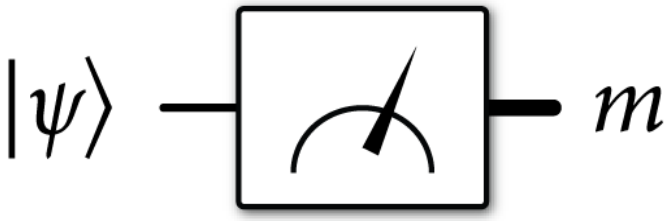

Figure 3.5: The above figure depicts our diagram of a quantum measurement. Thin lines denote quantum information and thick lines denote classical information. The result of the measurement is to output a classical variable m according to a probability distribution governed by the Born rule of the quantum theory.

$|\psi\rangle$ in (3.2). Nature prevents us from learning anything about the probability amplitudes α and β if we have only one quantum measurement that we can perform. Nature only allows us to measure *observables*. Observables are physical variables such as the position or momentum of a particle. In the quantum theory, we represent observables as Hermitian operators because their eigenvalues are real numbers and every measuring device outputs a real number. Examples of qubit observables that we can measure are the Pauli operators X, Y, and Z.

Suppose we measure the Z operator. This measurement is called a "measurement in the computational basis" or a "measurement of the Z observable" because we are measuring the eigenvalues of the Z operator. The measurement postulate of the quantum theory, also known as the *Born rule*, states that the system "collapses" into the state $|0\rangle$ with probability $|\alpha|^2$ and collapses into the state $|1\rangle$ with probability $|\beta|^2$. That is, the resulting probabilities are the squares of the probability amplitudes. After the measurement, our measuring apparatus tells us whether the state collapsed into $|0\rangle$ or $|1\rangle$—it returns $+1$ if the resulting state is $|0\rangle$ and returns -1 if the resulting state is $|1\rangle$. These returned values are the eigenvalues of the Z operator. The measurement postulate is the aspect of the quantum theory that makes it probabilistic or "jumpy" and is part of the "strangeness" of the quantum theory. Figure 3.5 depicts the notation for a measurement that we will use in diagrams throughout this book.

What is the result if we measure the state $|\psi\rangle$ in the $+/-$ basis? Consider that we can represent $|\psi\rangle$ as a superposition of the $|+\rangle$ and $|-\rangle$ states, as given in (3.35). The measurement postulate then states that a measurement of the X operator gives the state $|+\rangle$ with probability $|\alpha+\beta|^2/2$ and the state $|-\rangle$ with probability $|\alpha-\beta|^2/2$. Quantum interference is now playing a role because the amplitudes α and β interfere with each other. So this effect plays an important role in quantum information theory.

In some cases, the basis states $|0\rangle$ and $|1\rangle$ may not represent the spin states of an electron, but may represent the *location* of an electron. So, a way to interpret this measurement postulate is that the electron "jumps into" one location or another depending on the outcome of the measurement. But what is the state of the electron before the measurement? We will just say in this book that it is in a superposed, indefinite, or unsharp state, rather than trying to pin down a philosophical interpretation. Some might say that the electron is in

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Quantum State	**Probability of $\|+\rangle$**	**Probability of $\|-\rangle$**
Superposition state	$\|\alpha+\beta\|^2/2$	$\|\alpha-\beta\|^2/2$
Probabilistic description	$1/2$	$1/2$

Table 3.1: The above table summarizes the differences in probabilities for a quantum state in a superposition $\alpha|0\rangle + \beta|1\rangle$ and a classical state that is a probabilistic mixture of $|0\rangle$ and $|1\rangle$.

"two different locations at the same time."

Also, we should stress that we cannot interpret this measurement postulate as meaning that the state is in $|0\rangle$ or $|1\rangle$ with respective probabilities $|\alpha|^2$ and $|\beta|^2$ before the measurement occurs, because this latter scenario is completely classical. The superposition state $\alpha|0\rangle+\beta|1\rangle$ gives fundamentally different behavior from the probabilistic description of a state that is in $|0\rangle$ or $|1\rangle$ with respective probabilities $|\alpha|^2$ and $|\beta|^2$. Suppose that we have the two different descriptions of a state (superposition and probabilistic) and measure the Z operator. We get the same result for both cases—the resulting state is $|0\rangle$ or $|1\rangle$ with respective probabilities $|\alpha|^2$ and $|\beta|^2$.

But now suppose that we measure the X operator. The superposed state gives the result from before—we get the state $|+\rangle$ with probability $|\alpha+\beta|^2/2$ and the state $|-\rangle$ with probability $|\alpha-\beta|^2/2$. The probabilistic description gives a much different result. Suppose that the state is $|0\rangle$. We know that $|0\rangle$ is a uniform superposition of $|+\rangle$ and $|-\rangle$:

$$|0\rangle = \frac{|+\rangle + |-\rangle}{\sqrt{2}}. \tag{3.92}$$

So the state collapses to $|+\rangle$ or $|-\rangle$ with equal probability in this case. If the state is $|1\rangle$, then it collapses again to $|+\rangle$ or $|-\rangle$ with equal probabilities. Summing up these probabilities, it follows that a measurement of the X operator gives the state $|+\rangle$ with probability $\left(|\alpha|^2 + |\beta|^2\right)/2 = 1/2$ and gives the state $|-\rangle$ with the same probability. These results are fundamentally different from those where the state is the superposition state $|\psi\rangle$, and experiment after experiment supports the predictions of the quantum theory. Table 3.1 summarizes the results described in the above paragraph.

Now we consider a "Stern-Gerlach" like argument to illustrate another example of fundamental quantum behavior [101]. The Stern-Gerlach experiment was a crucial one for determining the "strange" behavior of quantum spin states. Suppose we prepare the state $|0\rangle$. If we measure this state in the Z basis, the result is that we always obtain the state $|0\rangle$ because it is a definite Z eigenstate. Suppose now that we measure the X operator. The state $|0\rangle$ is equivalent to a uniform superposition of $|+\rangle$ and $|-\rangle$. The measurement postulate then states that we get the state $|+\rangle$ or $|-\rangle$ with equal probability after performing this measurement. If we then measure the Z operator again, the result is completely random. The Z measurement result is $|0\rangle$ or $|1\rangle$ with equal probability if the result of the X measurement is $|+\rangle$ and the same distribution holds if the result of the X measurement is $|-\rangle$. This argument demonstrates that the measurement of the X operator throws off the measurement of the Z operator. The Stern-Gerlach experiment was one of the earliest to validate the predictions of the quantum theory.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

3.4.1 Probability, Expectation, and Variance of an Operator

We have an alternate, more formal way of stating the measurement postulate that turns out to be more useful for a general quantum system. Suppose that we are measuring the Z operator. The diagonal representation of this operator is

$$Z = |0\rangle\langle 0| - |1\rangle\langle 1|. \tag{3.93}$$

Consider the operator

$$\Pi_0 \equiv |0\rangle\langle 0|. \tag{3.94}$$

It is a projection operator because applying it twice has the same effect as applying it once: $\Pi_0^2 = \Pi_0$. It projects onto the subspace spanned by the single vector $|0\rangle$. A similar line of analysis applies to the projection operator

$$\Pi_1 \equiv |1\rangle\langle 1|. \tag{3.95}$$

So we can represent the Z operator as $\Pi_0 - \Pi_1$. Performing a measurement of the Z operator is equivalent to asking the question: Is the state $|0\rangle$ or $|1\rangle$? Consider the quantity $\langle\psi|\Pi_0|\psi\rangle$:

$$\langle\psi|\Pi_0|\psi\rangle = \langle\psi|0\rangle\langle 0|\psi\rangle = \alpha^*\alpha = |\alpha|^2. \tag{3.96}$$

A similar analysis demonstrates that

$$\langle\psi|\Pi_1|\psi\rangle = |\beta|^2. \tag{3.97}$$

These two quantities then give the probability that the state collapses to $|0\rangle$ or $|1\rangle$.

A more general way of expressing a measurement of the Z basis is to say that we have a set $\{\Pi_i\}_{i\in\{0,1\}}$ of measurement operators that determine the outcome probabilities. These measurement operators also determine the state that results after the measurement. If the measurement result is $+1$, then the resulting state is

$$\frac{\Pi_0|\psi\rangle}{\sqrt{\langle\psi|\Pi_0|\psi\rangle}} = |0\rangle, \tag{3.98}$$

where we implicitly ignore the irrelevant global phase factor $\frac{\alpha}{|\alpha|}$. If the measurement result is -1, then the resulting state is

$$\frac{\Pi_1|\psi\rangle}{\sqrt{\langle\psi|\Pi_1|\psi\rangle}} = |1\rangle, \tag{3.99}$$

where we again implicitly ignore the irrelevant global phase factor $\frac{\beta}{|\beta|}$. Dividing by $\sqrt{\langle\psi|\Pi_i|\psi\rangle}$ for $i = 0, 1$ ensures that the state resulting after measurement corresponds to a physical state that has unit norm.

We can also measure any orthonormal basis in this way—this type of projective measurement is called a *von Neumann measurement*. For any orthonormal basis $\{|\phi_i\rangle\}_{i\in\{0,1\}}$, the measurement operators are $\{|\phi_i\rangle\langle\phi_i|\}_{i\in\{0,1\}}$, and the state collapses to $|\phi_i\rangle\langle\phi_i|\psi\rangle/|\langle\phi_i|\psi\rangle|$ with probability $\langle\psi|\phi_i\rangle\langle\phi_i|\psi\rangle = |\langle\phi_i|\psi\rangle|^2$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 3.4.1 Determine the set of measurement operators corresponding to a measurement of the X observable.

We might want to determine the expected measurement result when measuring the Z operator. The probability of getting the $+1$ value corresponding to the $|0\rangle$ state is $|\alpha|^2$ and the probability of getting the -1 value corresponding to the -1 eigenstate is $|\beta|^2$. Standard probability theory then gives us a way to calculate the expected value of a measurement of the Z operator when the state is $|\psi\rangle$:

$$\mathbb{E}[Z] = |\alpha|^2(1) + |\beta|^2(-1) \tag{3.100}$$
$$= |\alpha|^2 - |\beta|^2. \tag{3.101}$$

We can formulate an alternate way to write this expectation, by making use of the Dirac notation:

$$\mathbb{E}[Z] = |\alpha|^2(1) + |\beta|^2(-1) \tag{3.102}$$
$$= \langle\psi|\Pi_0|\psi\rangle + \langle\psi|\Pi_1|\psi\rangle(-1) \tag{3.103}$$
$$= \langle\psi|\Pi_0 - \Pi_1|\psi\rangle \tag{3.104}$$
$$= \langle\psi|Z|\psi\rangle \tag{3.105}$$

It is common for physicists to denote the expectation as

$$\langle Z\rangle \equiv \langle\psi|Z|\psi\rangle, \tag{3.106}$$

when it is understood that the expectation is with respect to the state $|\psi\rangle$. This type of expression is a general one and the next exercise asks you to show that it works for the X and Y operators as well.

Exercise 3.4.2 Show that the expressions $\langle\psi|X|\psi\rangle$ and $\langle\psi|Y|\psi\rangle$ give the respective expectations $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ when measuring the state $|\psi\rangle$ in the respective X and Y basis.

We also might want to determine the variance of the measurement of the Z operator. Standard probability theory again gives that

$$\text{Var}[Z] = \mathbb{E}\left[Z^2\right] - \mathbb{E}[Z]^2. \tag{3.107}$$

Physicists denote the standard deviation of the measurement of the Z operator as

$$\Delta Z \equiv \left\langle (Z - \langle Z\rangle)^2 \right\rangle^{1/2}, \tag{3.108}$$

and thus the variance is equal to $(\Delta Z)^2$. Physicists often refer to ΔZ as the uncertainty of the observable Z when the state is $|\psi\rangle$.

In order to calculate the variance $\text{Var}[Z]$, we really just need the second moment $\mathbb{E}[Z^2]$ because we already have the expecation $\mathbb{E}[Z]$:

$$\mathbb{E}\left[Z^2\right] = |\alpha|^2(1)^2 + |\beta|^2(-1)^2 \tag{3.109}$$
$$= |\alpha|^2 + |\beta|^2. \tag{3.110}$$

We can again calculate this quantity with the Dirac notation. The quantity $\langle\psi|Z^2|\psi\rangle$ is the same as $\mathbb{E}[Z^2]$ and the next exercise asks you for a proof.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 3.4.3 Show that $\mathbb{E}[X^2] = \langle\psi|X^2|\psi\rangle$, $\mathbb{E}[Y^2] = \langle\psi|Y^2|\psi\rangle$, and $\mathbb{E}[Z^2] = \langle\psi|Z^2|\psi\rangle$.

3.4.2 The Uncertainty Principle

The uncertainty principle is a fundamental aspect of the quantum theory. In the case of qubits, one instance of the uncertainty principle gives a lower bound on the product of the uncertainty of the Z operator with the uncertainty of the X operator:

$$\Delta Z \Delta X \geq \frac{1}{2}|\langle\psi|[X, Z]|\psi\rangle|. \tag{3.111}$$

We can prove this principle using the postulates of the quantum theory. Let us define the operators $Z_0 \equiv Z - \langle Z\rangle$ and $X_0 \equiv X - \langle X\rangle$. First, consider that

$$\begin{aligned} \Delta Z \Delta X &= \langle\psi|Z_0^2|\psi\rangle^{1/2}\langle\psi|X_0^2|\psi\rangle^{1/2} && (3.112)\\ &\geq |\langle\psi|Z_0 X_0|\psi\rangle| && (3.113)\end{aligned}$$

The above step follows by applying the Cauchy-Schwarz inequality to the vectors $X_0|\psi\rangle$ and $Z_0|\psi\rangle$. For any operator A, we define its real part $\mathrm{Re}\{A\}$ as

$$\mathrm{Re}\{A\} \equiv \frac{A + A^\dagger}{2}, \tag{3.114}$$

and its imaginary part $\mathrm{Im}\{A\}$ as

$$\mathrm{Im}\{A\} \equiv \frac{A - A^\dagger}{2i}, \tag{3.115}$$

so that

$$A = \mathrm{Re}\{A\} + i\,\mathrm{Im}\{A\}. \tag{3.116}$$

So the real and imaginary parts of the operator $Z_0 X_0$ are

$$\begin{aligned} \mathrm{Re}\{Z_0 X_0\} &= \frac{Z_0 X_0 + X_0 Z_0}{2} \equiv \frac{\{Z_0, X_0\}}{2} && (3.117)\\ \mathrm{Im}\{Z_0 X_0\} &= \frac{Z_0 X_0 - X_0 Z_0}{2i} \equiv \frac{[Z_0, X_0]}{2i} && (3.118)\end{aligned}$$

where $\{Z_0, X_0\}$ is the anticommutator of Z_0 and X_0 and $[Z_0, X_0]$ is the commutator of the two operators. We can then express the quantity $|\langle\psi|Z_0 X_0|\psi\rangle|$ in terms of the real and imaginary parts of $Z_0 X_0$:

$$\begin{aligned} |\langle\psi|Z_0 X_0|\psi\rangle| &= |\langle\psi|\,\mathrm{Re}\{Z_0 X_0\}|\psi\rangle + i\langle\psi|\,\mathrm{Im}\{Z_0 X_0\}|\psi\rangle| && (3.119)\\ &\geq |\langle\psi|\,\mathrm{Im}\{Z_0 X_0\}|\psi\rangle| && (3.120)\\ &= |\langle\psi|[Z_0, X_0]|\psi\rangle|/2 && (3.121)\\ &= |\langle\psi|[Z, X]|\psi\rangle|/2 && (3.122)\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first equality follows by substitution, the first inequality follows because the magnitude of any complex number is greater than the magnitude of its imaginary part, the second equality follows by substitution with (3.118), and the third equality follows by the result of Exercise 3.4.4 below.

The commutator of the operators Z and X arises in the lower bound, and thus, the non-commutativity of the operators Z and X is the fundamental reason that there is an uncertainty principle for them. Also, there is no uncertainty principle for any two operators that commute with each other.

Exercise 3.4.4 Show that $[Z_0, X_0] = [Z, X]$ and that $[Z, X] = -2iY$.

Exercise 3.4.5 The uncertainty principle in (3.111) has the property that the lower bound has a dependence on the state $|\psi\rangle$. Find a state $|\psi\rangle$ for which the lower bound on the uncertainty product $\Delta X \Delta Z$ vanishes.[4]

3.5 Composite Quantum Systems

A single physical qubit is an interesting physical system that exhibits uniquely quantum phenomena, but it is not particularly useful on its own (just as a single classical bit is not particularly useful for classical communication or computation). We can only perform interesting quantum information processing tasks when we combine qubits together. Therefore, we should have a way for describing their behavior when they combine to form a composite quantum system.

Consider two classical bits c_0 and c_1. In order to describe bit operations on the pair of cbits, we write them as an ordered pair (c_1, c_0). The space of all possible bit values is the Cartesian product $\mathbb{Z}_2 \times \mathbb{Z}_2$ of two copies of the set $\mathbb{Z}_2 \equiv \{0, 1\}$:

$$\mathbb{Z}_2 \times \mathbb{Z}_2 \equiv \{(0,0), (0,1), (1,0), (1,1)\}. \tag{3.123}$$

Typically, we make the abbreviation $c_1 c_0 \equiv (c_1, c_0)$ when representing cbit states.

We can represent the state of two cbits with particular states of qubits. For example, we can represent the two-cbit state 00 with the following mapping:

$$00 \rightarrow |0\rangle|0\rangle. \tag{3.124}$$

Many times, we make the abbreviation $|00\rangle \equiv |0\rangle|0\rangle$ when representing two-cbit states with qubits. In general, any two-cbit state $c_1 c_0$ has the following representation as a two-qubit state:

$$c_1 c_0 \rightarrow |c_1 c_0\rangle. \tag{3.125}$$

[4]Do not be alarmed by the result of this exercise! The usual formulation of the uncertainty principle only gives a lower bound on the uncertainty product. This lower bound never vanishes for the case of position and momentum because the commutator of these two observables is equal to the identity operator multiplied by i, but it can vanish for the operators given in the exercise.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above qubit states are not the only possible states that can occur in the quantum theory. By the superposition principle, any possible linear combination of the set of two-cbit states is a possible two-qubit state:

$$|\xi\rangle \equiv \alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle. \tag{3.126}$$

The unit-norm condition $|\alpha|^2 + |\beta|^2 + |\gamma|^2 + |\delta|^2 = 1$ again must hold for the two-qubit state to correspond to a physical quantum state. It is now clear that the Cartesian product is not sufficient for representing two-qubit quantum states because it does not allow for linear combinations of states (just as the mathematics of Boolean algebra is not sufficient to represent single-qubit states).

We again turn to linear algebra to determine a representation that suffices. The *tensor product* is the mathematical operation that gives a sufficient representation of two-qubit quantum states. Suppose we have two two-dimensional vectors:

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix}, \qquad \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}. \tag{3.127}$$

The tensor product of these two vectors is

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix} \otimes \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \equiv \begin{bmatrix} a_1 \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \\ b_1 \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} a_1 a_2 \\ a_1 b_2 \\ b_1 a_2 \\ b_1 b_2 \end{bmatrix}. \tag{3.128}$$

Recall, from (3.4-3.5), the vector representation of the single-qubit states $|0\rangle$ and $|1\rangle$. Using these vector representations and the above definition of the tensor product, the two-qubit basis states have the following vector representations:

$$|00\rangle = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad |01\rangle = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad |10\rangle = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad |11\rangle = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{3.129}$$

A simple way to remember these representations is that the bits inside the ket index the element equal to one in the vector. For example, the vector representation of $|01\rangle$ has a one as its second element because 01 is the second index for the two-bit strings. The vector representation of the superposition state in (3.126) is

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix}. \tag{3.130}$$

There are actually many different ways that we can write two-qubit states, and we go through all of these right now. Physicists have developed many shorthands, and it is important to know each of these because they often appear in the literature (we even use different

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

notations depending on the context). We may use any of the following two-qubit notations if the two qubits are local to one party and only one party is involved in a protocol:

$$\alpha|0\rangle \otimes |0\rangle + \beta|0\rangle \otimes |1\rangle + \gamma|1\rangle \otimes |0\rangle + \delta|1\rangle \otimes |1\rangle, \tag{3.131}$$
$$\alpha|0\rangle|0\rangle + \beta|0\rangle|1\rangle + \gamma|1\rangle|0\rangle + \delta|1\rangle|1\rangle, \tag{3.132}$$
$$\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle. \tag{3.133}$$

We can put labels on the qubits if two or more parties are involved in the protocol:

$$\alpha|0\rangle^A \otimes |0\rangle^B + \beta|0\rangle^A \otimes |1\rangle^B + \gamma|1\rangle^A \otimes |0\rangle^B + \delta|1\rangle^A \otimes |1\rangle^B, \tag{3.134}$$
$$\alpha|0\rangle^A|0\rangle^B + \beta|0\rangle^A|1\rangle^B + \gamma|1\rangle^A|0\rangle^B + \delta|1\rangle^A|1\rangle^B, \tag{3.135}$$
$$\alpha|00\rangle^{AB} + \beta|01\rangle^{AB} + \gamma|10\rangle^{AB} + \delta|11\rangle^{AB}. \tag{3.136}$$

This second scenario is different from the first scenario because two spatially separated parties share the two-qubit state. If the state has quantum correlations, then it can be valuable as a communication resource. We go into more detail on this topic in Section 3.5.6 on *entanglement*.

3.5.1 Evolution of Composite Systems

The postulate on unitary evolution extends to the two-qubit scenario as well. First, let us establish that the tensor product $A \otimes B$ of two operators A and B is

$$A \otimes B \equiv \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \tag{3.137}$$
$$\equiv \begin{bmatrix} a_{11}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} & a_{12}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ a_{21}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} & a_{22}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \end{bmatrix} \tag{3.138}$$
$$= \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{bmatrix}. \tag{3.139}$$

Consider the two-qubit state in (3.126). We can perform a NOT gate on the first qubit so that it changes to

$$\alpha|10\rangle + \beta|11\rangle + \gamma|00\rangle + \delta|01\rangle. \tag{3.140}$$

We can likewise flip its second qubit:

$$\alpha|01\rangle + \beta|00\rangle + \gamma|11\rangle + \delta|10\rangle, \tag{3.141}$$

or flip both at the same time:

$$\alpha|11\rangle + \beta|10\rangle + \gamma|01\rangle + \delta|00\rangle. \tag{3.142}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

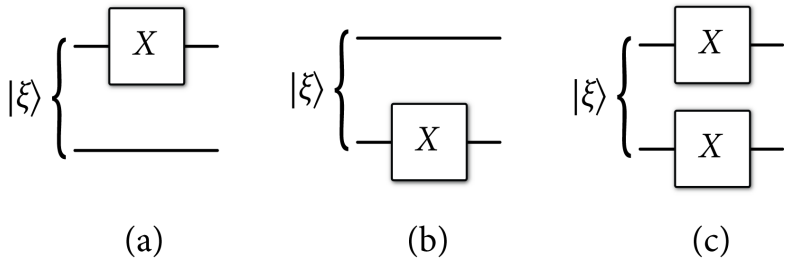

Figure 3.6: The above figure depicts circuits for the example two-qubit unitaries $X_1 I_2$, $I_1 X_2$, and $X_1 X_2$.

Figure 3.6 depicts quantum circuit representations of these operations. These are all reversible operations because applying them again gives the original state in (3.126). In the first case, we did nothing to the second qubit, and in the second case, we did nothing to the first qubit. The identity operator acts on the qubits that have nothing happen to them.

Let us label the first qubit as "1" and the second qubit as "2." We can then label the operator for the first operation as $X_1 I_2$ because this operator flips the first qubit and does nothing (applies the identity) to the second qubit. We can also label the operators for the second and third operations respectively as $I_1 X_2$ and $X_1 X_2$. The matrix representation of the operator $X_1 I_2$ is the tensor product of the matrix representation of X with the matrix representation of I—this relation similarly holds for the operators $I_1 X_2$ and $X_1 X_2$. We show that it holds for the operator $X_1 I_2$ and ask you to verify the other two cases. We can use the two-qubit computational basis to get a matrix representation for the two-qubit operator $X_1 I_2$:

$$\begin{bmatrix} \langle 00|X_1 I_2|00\rangle & \langle 00|X_1 I_2|01\rangle & \langle 00|X_1 I_2|10\rangle & \langle 00|X_1 I_2|11\rangle \\ \langle 01|X_1 I_2|00\rangle & \langle 01|X_1 I_2|01\rangle & \langle 01|X_1 I_2|10\rangle & \langle 01|X_1 I_2|11\rangle \\ \langle 10|X_1 I_2|00\rangle & \langle 10|X_1 I_2|01\rangle & \langle 10|X_1 I_2|10\rangle & \langle 10|X_1 I_2|11\rangle \\ \langle 11|X_1 I_2|00\rangle & \langle 11|X_1 I_2|01\rangle & \langle 11|X_1 I_2|10\rangle & \langle 11|X_1 I_2|11\rangle \end{bmatrix}$$
$$= \begin{bmatrix} \langle 00|10\rangle & \langle 00|11\rangle & \langle 00|00\rangle & \langle 00|01\rangle \\ \langle 01|10\rangle & \langle 01|11\rangle & \langle 01|00\rangle & \langle 01|01\rangle \\ \langle 10|10\rangle & \langle 10|11\rangle & \langle 10|00\rangle & \langle 10|01\rangle \\ \langle 11|10\rangle & \langle 11|11\rangle & \langle 11|00\rangle & \langle 11|01\rangle \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad (3.143)$$

This last matrix is equal to the tensor product $X \otimes I$ by inspecting the definition of the tensor product for matrices in (3.137).

Exercise 3.5.1 Show that the matrix representation of the operator $I_1 X_2$ is equal to the tensor product $I \otimes X$. Show the same for $X_1 X_2$ and $X \otimes X$.

3.5.2 Probability Amplitudes for Composite Systems

We relied on the orthogonality of the two-qubit computational basis states for evaluating amplitudes such as $\langle 00|10\rangle$ or $\langle 00|00\rangle$ in the above matrix representation. It turns out that

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

there is another way to evaluate these amplitudes that relies only on the orthogonality of the single-qubit computational basis states. Suppose that we have four single-qubit states $|\phi_0\rangle$, $|\phi_1\rangle$, $|\psi_0\rangle$, $|\psi_1\rangle$, and we make the following two-qubit states from them:

$$|\phi_0\rangle \otimes |\psi_0\rangle, \tag{3.144}$$
$$|\phi_1\rangle \otimes |\psi_1\rangle. \tag{3.145}$$

We may represent these states equally well as follows:

$$|\phi_0, \psi_0\rangle, \tag{3.146}$$
$$|\phi_1, \psi_1\rangle. \tag{3.147}$$

because the Dirac notation is versatile (virtually anything can go inside a ket as long as its meaning is not ambiguous). The bra $\langle\phi_1, \psi_1|$ is dual to the ket $|\phi_1, \psi_1\rangle$, and we can use it to calculate the following amplitude:

$$\langle\phi_1, \psi_1|\phi_0, \psi_0\rangle. \tag{3.148}$$

This amplitude is equivalent to the multiplication of the single-qubit amplitudes:

$$\langle\phi_1, \psi_1|\phi_0, \psi_0\rangle = \langle\phi_1|\phi_0\rangle\langle\psi_1|\psi_0\rangle. \tag{3.149}$$

Exercise 3.5.2 Verify that the amplitudes $\{\langle ij|kl\rangle\}_{i,j,k,l\in\{0,1\}}$ are respectively equal to the amplitudes $\{\langle i|k\rangle\langle j|l\rangle\}_{i,j,k,l\in\{0,1\}}$. By linearity, this exercise justifies the relation in (3.149) (at least for two-qubit states).

3.5.3 Controlled Gates

An important two-qubit unitary evolution is the controlled-NOT (CNOT) gate. We consider its classical version first. The classical gate acts on two cbits. It does nothing if the first bit is equal to zero, and flips the second bit if the first bit is equal to one:

$$00 \rightarrow 00, \qquad 01 \rightarrow 01, \qquad 10 \rightarrow 11, \qquad 11 \rightarrow 10. \tag{3.150}$$

We turn this gate into a quantum gate[5] by demanding that it act in the same way on the two-qubit computational basis states:

$$|00\rangle \rightarrow |00\rangle, \qquad |01\rangle \rightarrow |01\rangle, \qquad |10\rangle \rightarrow |11\rangle, \qquad |11\rangle \rightarrow |10\rangle. \tag{3.151}$$

This behavior carries over to superposition states as well:

$$\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle \quad \underrightarrow{\text{CNOT}} \quad \alpha|00\rangle + \beta|01\rangle + \gamma|11\rangle + \delta|10\rangle. \tag{3.152}$$

[5]There are other terms for the action of turning a classical operation into a quantum one. Some examples are "making it coherent," "coherifying," or the quantum gate is a "coherification" of the classical one. The term "coherify" is not a proper English word, but we will use it regardless at certain points.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

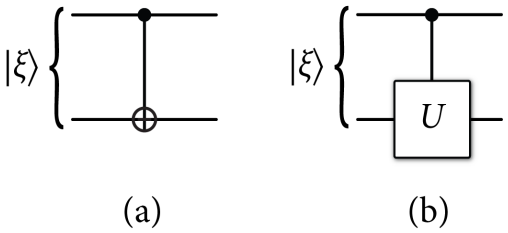

Figure 3.7: The above figure depicts the circuit diagrams that we use for (a) a CNOT gate and (b) a controlled-U gate.

A useful operator representation of the CNOT gate is

$$\text{CNOT} \equiv |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes X. \tag{3.153}$$

The above representation truly captures the coherent quantum nature of the CNOT gate. In the classical CNOT gate, we can say that it is a conditional gate, in the sense that the gate applies to the second bit conditional on the value of the first bit. In the quantum CNOT gate, the second operation is *controlled* on the basis state of the first qubit (hence the choice of the name "controlled-NOT"). That is, the gate always applies the second operation regardless of the actual qubit state on which it acts.

A controlled-U gate is similar to the CNOT gate in (3.153). It simply applies the unitary U (assumed to be a single-qubit unitary) to the second qubit, controlled on the first qubit:

$$\text{Controlled-}U \equiv |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes U. \tag{3.154}$$

The control qubit can be controlled with respect to any orthonormal basis $\{|\phi_0\rangle, |\phi_1\rangle\}$:

$$|\phi_0\rangle\langle\phi_0| \otimes I + |\phi_1\rangle\langle\phi_1| \otimes U. \tag{3.155}$$

Figure 3.7 depicts the circuit diagrams for a controlled-NOT and controlled-U operation.

Exercise 3.5.3 Verify that the matrix representation of the CNOT gate in the computational basis is

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{3.156}$$

Exercise 3.5.4 Consider applying Hadamards to the first and second qubits before and after a CNOT acts on them. Show that this gate is equivalent to a CNOT in the $+/-$ basis (recall that the Z operator flips the $+/-$ basis):

$$H_1 H_2 \text{ CNOT } H_1 H_2 = |+\rangle\langle +| \otimes I + |-\rangle\langle -| \otimes Z. \tag{3.157}$$

Example 3.5.1 Show that two CNOT gates with the same control qubit commute.

Exercise 3.5.5 Show that two CNOT gates with the same target qubit commute.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

3.5.4 The No Cloning Theorem

The no cloning theorem is one of the simplest results in the quantum theory, yet it has some of the most profound consequences. It states that it is impossible to build a *universal copier* of quantum states. A universal copier would be a device that could copy any arbitrary quantum state that is input to it. It may be surprising at first to hear that copying quantum information is impossible because copying classical information is ubiquitous.

We give a simple proof for the no-cloning theorem. Suppose for a contradiction that there is a two-qubit unitary operator U acting as a universal copier of quantum information. That is, if we input an arbitrary state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ as the first qubit and input an ancilla qubit $|0\rangle$ as the second qubit, it "writes" the first qubit to the second qubit slot as follows:

$$\begin{aligned} U|\psi\rangle|0\rangle &= |\psi\rangle|\psi\rangle & (3.158)\\ &= (\alpha|0\rangle + \beta|1\rangle)(\alpha|0\rangle + \beta|1\rangle) & (3.159)\\ &= \alpha^2|0\rangle|0\rangle + \alpha\beta|0\rangle|1\rangle + \alpha\beta|1\rangle|0\rangle + \beta^2|1\rangle|1\rangle. & (3.160)\end{aligned}$$

The copier is universal, meaning that it copies an arbitrary state. In particular, it also copies the states $|0\rangle$ and $|1\rangle$:

$$\begin{aligned} U|0\rangle|0\rangle &= |0\rangle|0\rangle, & (3.161)\\ U|1\rangle|0\rangle &= |1\rangle|1\rangle. & (3.162)\end{aligned}$$

Linearity of the quantum theory then implies that the unitary operator acts on a superposition $\alpha|0\rangle + \beta|1\rangle$ as follows:

$$U(\alpha|0\rangle + \beta|1\rangle)|0\rangle = \alpha|0\rangle|0\rangle + \beta|1\rangle|1\rangle. \quad (3.163)$$

The result in (3.160) contradicts the result in (3.163) because these two expressions do not have to be equal for all α and β:

$$\exists \alpha, \beta : \alpha^2|0\rangle|0\rangle + \alpha\beta|0\rangle|1\rangle + \alpha\beta|1\rangle|0\rangle + \beta^2|1\rangle|1\rangle \neq \alpha|0\rangle|0\rangle + \beta|1\rangle|1\rangle. \quad (3.164)$$

Thus, unitarity in the quantum theory contradicts the existence of a universal quantum copier.

We would like to stress that this proof does not mean that it is impossible to copy certain quantum states—it only implies the impossibility of a *universal* copier. Another proof of the no-cloning theorem gives insight into the type of states that we can copy. Let us again suppose that a universal copier U exists. Consider two arbitrary states $|\psi\rangle$ and $|\phi\rangle$. If a universal copier U exists, then it performs the following copying operation for both states:

$$\begin{aligned} U|\psi\rangle|0\rangle &= |\psi\rangle|\psi\rangle, & (3.165)\\ U|\phi\rangle|0\rangle &= |\phi\rangle|\phi\rangle. & (3.166)\end{aligned}$$

Consider the probability amplitude $\langle\psi|\langle\psi||\phi\rangle|\phi\rangle$:

$$\langle\psi|\langle\psi||\phi\rangle|\phi\rangle = \langle\psi|\phi\rangle\langle\psi|\phi\rangle = \langle\psi|\phi\rangle^2. \quad (3.167)$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The following relation for $\langle\psi|\langle\psi||\phi\rangle|\phi\rangle$ holds as well by using the results in (3.165) and the unitarity property $U^\dagger U = I$:

$$\begin{aligned}\langle\psi|\langle\psi||\phi\rangle|\phi\rangle &= \langle\psi|\langle 0|U^\dagger U|\phi\rangle|0\rangle && (3.168)\\ &= \langle\psi|\langle 0||\phi\rangle|0\rangle && (3.169)\\ &= \langle\psi|\phi\rangle\langle 0|0\rangle && (3.170)\\ &= \langle\psi|\phi\rangle. && (3.171)\end{aligned}$$

It then holds that

$$\langle\psi|\langle\psi||\phi\rangle|\phi\rangle = \langle\psi|\phi\rangle^2 = \langle\psi|\phi\rangle, \tag{3.172}$$

by employing the above two results. The relation $\langle\psi|\phi\rangle^2 = \langle\psi|\phi\rangle$ holds for exactly two cases, $\langle\psi|\phi\rangle = 1$ and $\langle\psi|\phi\rangle = 0$. The first case holds only when the two states are the same state and the second case holds when the two states are orthogonal to each other. Thus, it is impossible to copy quantum information in any other case because we would again contradict unitarity.

The no-cloning theorem has several applications in quantum information processing. First, it underlies the security of the quantum key distribution protocol because it ensures that an attacker cannot copy the quantum states that two parties use to establish a secret key. It finds application in quantum Shannon theory because we can use it to reason about the quantum capacity of a certain quantum channel known as the erasure channel. We will return to this point in Chapter 23.

Exercise 3.5.6 Suppose that two states $|\psi\rangle$ and $|\psi^\perp\rangle$ are orthogonal:

$$\langle\psi|\psi^\perp\rangle = 0. \tag{3.173}$$

Construct a two-qubit unitary that can copy the states, i.e., find a unitary U that acts as follows:

$$\begin{aligned}U|\psi\rangle|0\rangle &= |\psi\rangle|\psi\rangle, && (3.174)\\ U|\psi^\perp\rangle|0\rangle &= |\psi^\perp\rangle|\psi^\perp\rangle. && (3.175)\end{aligned}$$

Exercise 3.5.7 (No-deletion theorem) Related to the no-cloning theorem, there is a no-deletion theorem. Suppose that two copies of a quantum state $|\psi\rangle$ are available, and the goal is to delete one of these states by a unitary interaction. That is, there should exist a universal quantum deleter U that has the following action on the two copies of $|\psi\rangle$ and an ancilla state $|A\rangle$, regardless of the input state $|\psi\rangle$:

$$U|\psi\rangle|\psi\rangle|A\rangle = |\psi\rangle|0\rangle|A'\rangle. \tag{3.176}$$

Show that this is impossible.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

3.5.5 Measurement of Composite Systems

The measurement postulate also extends to composite quantum systems. Suppose again that we have the two-qubit quantum state in (3.126). By a straightforward analogy with the single-qubit case, we can determine the following amplitudes:

$$\langle 00|\xi\rangle = \alpha, \quad \langle 01|\xi\rangle = \beta, \quad \langle 10|\xi\rangle = \gamma, \quad \langle 11|\xi\rangle = \delta. \tag{3.177}$$

We can also define the following projection operators

$$\Pi_{00} \equiv |00\rangle\langle 00|, \quad \Pi_{01} \equiv |01\rangle\langle 01|, \quad \Pi_{10} \equiv |10\rangle\langle 10|, \quad \Pi_{00} \equiv |11\rangle\langle 11|, \tag{3.178}$$

and apply the Born rule to determine the probabilities for each result:

$$\langle \xi|\Pi_{00}|\xi\rangle = |\alpha|^2, \quad \langle \xi|\Pi_{01}|\xi\rangle = |\beta|^2, \quad \langle \xi|\Pi_{10}|\xi\rangle = |\gamma|^2, \quad \langle \xi|\Pi_{11}|\xi\rangle = |\delta|^2. \tag{3.179}$$

Suppose that we wish to perform a measurement of the Z operator on the first qubit only. What is the set of projection operators that describes this measurement? The answer is similar to what we found for the evolution of a composite system. We apply the identity operator to the second qubit because no measurement occurs on it. Thus, the set of measurement operators is

$$\{\Pi_0 \otimes I, \Pi_1 \otimes I\}, \tag{3.180}$$

where the definition of Π_0 and Π_1 is in (3.94-3.95). The state collapses to

$$\frac{(\Pi_0 \otimes I)|\xi\rangle}{\sqrt{\langle \xi|(\Pi_0 \otimes I)|\xi\rangle}} = \frac{\alpha|00\rangle + \beta|01\rangle}{\sqrt{|\alpha|^2 + |\beta|^2}}, \tag{3.181}$$

with probability $\langle \xi|(\Pi_0 \otimes I)|\xi\rangle = |\alpha|^2 + |\beta|^2$, and collapses to

$$\frac{(\Pi_1 \otimes I)|\xi\rangle}{\sqrt{\langle \xi|(\Pi_1 \otimes I)|\xi\rangle}} = \frac{\gamma|10\rangle + \delta|11\rangle}{\sqrt{|\gamma|^2 + |\delta|^2}}, \tag{3.182}$$

with probability $\langle \xi|(\Pi_1 \otimes I)|\xi\rangle = |\gamma|^2 + |\delta|^2$. The divisions by $\sqrt{\langle \xi|(\Pi_0 \otimes I)|\xi\rangle}$ and $\sqrt{\langle \xi|(\Pi_1 \otimes I)|\xi\rangle}$ again ensure that the resulting state is a normalized, physical state.

3.5.6 Entanglement

Composite quantum systems give rise to the most uniquely quantum phenomenon: *entanglement*. Schrödinger first observed that two or more quantum systems can be entangled and coined the term after noticing some of the bizarre consequences of this phenomenon.[6]

[6]Schrodinger actually used the German word "Verschränkung" to describe the phenomenon, which literally translates as "little parts that, though far from one another, always keep the exact same distance from each other." The one-word English translation is "entanglement." Einstein described the "Verschränkung" as a "spukhafte Fernwirkung," most closely translated as "long-distance ghostly effect" or the more commonly stated "spooky action at a distance."

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We first consider a simple, unentangled state that two parties, Alice and Bob, may share, in order to see how an unentangled state contrasts with an entangled state. Suppose that they share the state

$$|0\rangle^A|0\rangle^B, \tag{3.183}$$

where Alice has the qubit in system A and Bob has the qubit in system B. Alice can definitely say that her qubit is in the state $|0\rangle^A$ and Bob can definitely say that his qubit is in the state $|0\rangle^B$. There is nothing really too strange about this scenario.

Now, consider the composite quantum state $|\Phi^+\rangle^{AB}$:

$$\left|\Phi^+\right\rangle^{AB} \equiv \frac{|0\rangle^A|0\rangle^B + |1\rangle^A|1\rangle^B}{\sqrt{2}}. \tag{3.184}$$

Alice again has possession of the first qubit in system A and Bob has possession of the second qubit in system B. But now, it is not clear from the above description how to determine the individual state of Alice or the individual state of Bob. The above state is really a uniform superposition of the joint state $|0\rangle^A|0\rangle^B$ and the joint state $|1\rangle^A|1\rangle^B$, and it is not possible to describe either Alice's or Bob's individual state in the noiseless quantum theory. We also cannot describe the entangled state $|\Phi^+\rangle^{AB}$ as a product state of the form $|\phi\rangle^A|\psi\rangle^B$. This criterion is one possible criterion to determine if a state is entangled, but we will see a more formal criterion in the coming section.

Exercise 3.5.8 Show that the entangled state $|\Phi^+\rangle^{AB}$ has the following representation in the $+/-$ basis:

$$\left|\Phi^+\right\rangle^{AB} = \frac{|+\rangle^A|+\rangle^B + |-\rangle^A|-\rangle^B}{\sqrt{2}}. \tag{3.185}$$

Figure 3.8 gives a graphical depiction of entanglement. We use this depiction often throughout this book. Alice and Bob must receive the entanglement in some way, and the diagram indicates that some source distributes the entangled pair to them. It indicates that Alice and Bob are spatially separated and they possess the entangled state after some time. If they share the entangled state in (3.184), we say that they share one bit of entanglement, or one *ebit*. The term "ebit" implies that there is some way to quantify entanglement and we will make this notion clear in Chapter 18.

Entanglement as a Resource

In this book, we are interested in the use of entanglement as a resource. Much of this book concerns the theory of quantum information processing resources and we have a standard notation for the theory of resources. Let us represent the resource of a shared ebit as

$$[qq], \tag{3.186}$$

meaning that the ebit is a noiseless, quantum resource shared between two parties. Square brackets indicate a noiseless resource, the letter q indicates a quantum resource, and the two copies of the letter q indicate a two-party resource.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

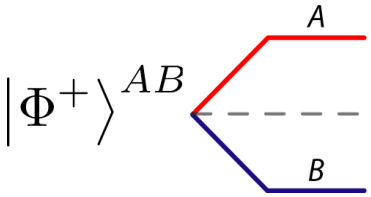

Figure 3.8: We use the above diagram to depict entanglement shared between two parties A and B. The diagram indicates that a source location creates the entanglement and distributes one system (the red system) to A and the other system (the blue system) to B. The standard unit of entanglement is the ebit in a Bell state $|\Phi^+\rangle \equiv (|00\rangle^{AB} + |11\rangle^{AB})/\sqrt{2}$.

Our first example of the use of entanglement is its role in generating *common randomness*. We define one bit of common randomness as the following probability distribution for two binary random variables X_A and X_B:

$$p_{X_A,X_B}(x_A, x_B) = \frac{1}{2}\delta(x_A, x_B), \tag{3.187}$$

where δ is the Kronecker delta function. Suppose Alice possesses random variable X_A and Bob possesses random variable X_B. Thus, with probability 1/2, they either both have a zero or they both have a one. We represent the resource of one bit of common randomness as

$$[cc], \tag{3.188}$$

indicating that a bit of common randomness is a noiseless, classical resource shared between two parties.

Now suppose that Alice and Bob share an ebit and they decide that they will each measure their qubits in the computational basis. Without loss of generality, suppose that Alice performs a measurement first. Thus, Alice performs a measurement of the Z^A operator, meaning that she measures $Z^A \otimes I^B$ (she cannot perform anything on Bob's qubit because they are spatially separated). The projection operators for this measurement are the same from (3.180) and just before Alice looks at her measurement result, we can describe the system as being in the following ensemble of states:

$$|0\rangle^A|0\rangle^B \text{ with probability } \frac{1}{2}, \tag{3.189}$$

$$|1\rangle^A|1\rangle^B \text{ with probability } \frac{1}{2}. \tag{3.190}$$

The interesting thing about the above ensemble is that Bob's result is already determined even before he measures, just after Alice looks at her result. Suppose that Alice knows the result of her measurement is $|0\rangle^A$. When Bob measures his system, he obtains the state $|0\rangle^B$ with probability one and *Alice knows that he has measured this result*. Additionally, Bob knows that Alice's state is $|0\rangle^A$ if he obtains $|0\rangle^B$. The same results hold if Alice knows that

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

the result of her measurement is $|1\rangle^A$. Thus, this protocol is a method for them to generate one bit of common randomness as defined in (3.187).

We can phrase the above protocol as the following *resource inequality*:

$$[qq] \geq [cc]. \tag{3.191}$$

The interpretation of the above resource inequality is *that there exists a protocol which generates the resource on the right by consuming the resource on the left*, and for this reason, the resource on the left is a stronger resource than the one on the right. The theory of resource inequalities plays a prominent role in this book and is a useful shorthand for expressing quantum protocols.

A natural question is to wonder if there exists a protocol to generate entanglement from common randomness. It is not possible to do so and the reason for this inequivalence of resources is another type of inequality (different from the resource inequality mentioned above), called a Bell's inequality. In short, Bell's theorem places an upper bound on the correlations present in any two classical systems. Entanglement violates this inequality, showing that it has no known classical equivalent. Thus, entanglement is a strictly stronger resource than common randomness and the resource inequality in (3.191) only holds in the given direction.

Common randomness is a resource in classical information theory, and may be useful in some scenarios, but it is actually a rather weak resource. Surely, generating common randomness is not the only use of entanglement. It turns out that we can construct far more exotic protocols such as the teleportation protocol or the super-dense coding protocol by combining the resource of entanglement with other resources. We discuss these protocols in Chapter 6.

Exercise 3.5.9 Use the representation of the ebit in Exercise 3.5.8 to show that Alice and Bob can measure the X operator to generate common randomness. This ability to obtain common randomness by both parties measuring in either the Z or X basis is the basis for an entanglement-based secret key distribution protocol.

Exercise 3.5.10 (Cloning implies signaling) Prove that if a universal quantum cloner were to exist, then it would be possible for Alice to signal to Bob faster than the speed of light by exploiting only the ebit state $|\Phi^+\rangle^{AB}$ shared between them and no communication. That is, show the existence of a protocol that would allow for this. (Hint: One possibility is for Alice to measure the X or Z Pauli operator locally on her share of the ebit, and then for Bob to exploit the universal quantum cloner. Consider the representation of the ebit in (3.184) and (3.185).)

Entanglement in the CHSH Game

One of the simplest means for demonstrating the power of entanglement is with a two-player game known as the CHSH game (after Clauser, Horne, Shimony, and Holt). We first present the rules of the game, and then we find an upper bound on the probability that players

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

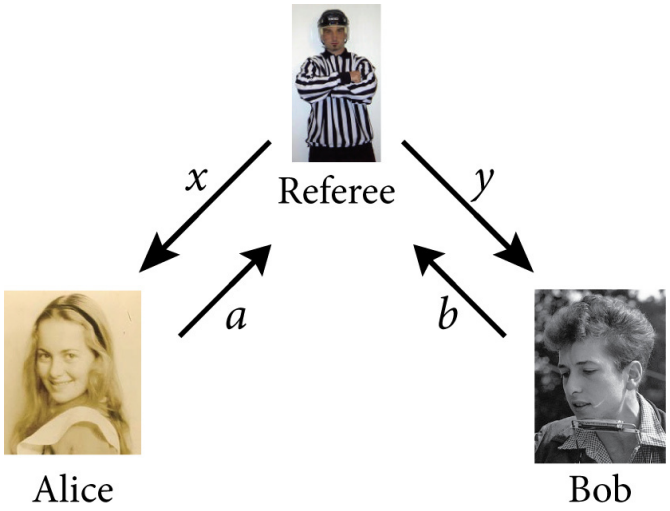

Figure 3.9: A depiction of the CHSH game. The referee distributes the bits x and y to Alice and Bob in the first round. In the second round, Alice and Bob return the bits a and b to the referee.

sharing classical correlations can win. We finally leave it as an exercise to show that players sharing a maximally entangled Bell state $|\Phi^+\rangle$ can have an approximately 10% higher chance of winning the game with a quantum strategy.

The players of the game are Alice and Bob. The game begins with a referee selecting two bits x and y uniformly at random. He then sends x to Alice and y to Bob. Alice and Bob are not allowed to communicate in any way at this point. Alice sends back to the referee a bit a, and Bob sends back a bit b. Since they are spatially separated, Alice's bit a can only depend on x, and similarly, Bob's bit b can only depend on y. The referee then determines if the AND of x and y is equal to the exclusive OR of a and b. If so, then Alice and Bob win the game. That is, the winning condition is

$$x \wedge y = a \oplus b. \tag{3.192}$$

Figure 3.9 depicts the CHSH game.

Before the game begins, Alice and Bob are allowed to coordinate on a strategy. A deterministic strategy would have Alice select a bit a_x conditioned on the bit x that she receives, and similarly, Bob would select a bit b_y conditioned on y. The following table presents the winning conditions for the four different values of x and y with this deterministic strategy:

x	y	$x \wedge y$	$= a_x \oplus b_y$
0	0	0	$= a_0 \oplus b_0$
0	1	0	$= a_0 \oplus b_1$
1	0	0	$= a_1 \oplus b_0$
1	1	1	$= a_1 \oplus b_1$

. (3.193)

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Though, we can observe that it is impossible for them to always win. If we add the entries in the column $x \wedge y$, the binary sum is equal to one, while if we add the entries in the column $= a_x \oplus b_y$, the binary sum is equal to zero. Thus, it is impossible for all of these equations to be satisfied. At most, only three out of four of them can be satisfied, so that the maximal winning probability with a classical deterministic strategy is at most 3/4. This upper bound also serves as an upper bound on the winning probability for the case in which they employ a randomized strategy coordinated by shared randomness—any such strategy would just be a convex combination of deterministic strategies. We can then see that a strategy for them to achieve this upper bound is for Alice and Bob always to return $a = 0$ and $b = 0$ no matter the values of x and y.

Interestingly, if Alice and Bob share a maximally entangled state, they can achieve a higher winning probability than if they share classical correlations only. This is one demonstration of the power of entanglement, and we leave it as an exercise to prove that the following quantum strategy achieves a winning probability of $\cos^2(\pi/8) \approx 0.85$ in the CHSH game.

Exercise 3.5.11 Suppose that Alice and Bob share a maximally entangled state $|\Phi^+\rangle$. Show that the following strategy has a winning probability of $\cos^2(\pi/8)$. If Alice receives $x = 0$ from the referee, then she performs a measurement of Pauli Z on her system and returns the outcome as a. If she receives $x = 1$, then she performs a measurement of Pauli X and returns the outcome as a. If Bob receives $y = 0$ from the referee, then he performs a measurement of $(X + Z)/\sqrt{2}$ on his system and returns the outcome as b. If Bob receives $y = 1$ from the referee, then he performs a measurement of $(Z - X)/\sqrt{2}$ and returns the outcome as b.

The Bell States

There are other useful entangled states besides the standard ebit. Suppose that Alice performs a Z^A operation on her half of the ebit $|\Phi^+\rangle^{AB}$. Then the resulting state is

$$\left|\Phi^-\right\rangle^{AB} \equiv \frac{1}{\sqrt{2}}\Big(|00\rangle^{AB} - |11\rangle^{AB}\Big). \tag{3.194}$$

Similarly, if Alice performs an X operator or a Y operator, the global state transforms to the following respective states (up to a global phase):

$$\left|\Psi^+\right\rangle^{AB} \equiv \frac{1}{\sqrt{2}}\Big(|01\rangle^{AB} + |10\rangle^{AB}\Big), \tag{3.195}$$

$$\left|\Psi^-\right\rangle^{AB} \equiv \frac{1}{\sqrt{2}}\Big(|01\rangle^{AB} - |10\rangle^{AB}\Big). \tag{3.196}$$

The states $|\Phi^+\rangle^{AB}$, $|\Phi^-\rangle^{AB}$, $|\Psi^+\rangle^{AB}$, and $|\Psi^-\rangle^{AB}$ are known as the *Bell states* and are the most important entangled states for a two-qubit system. They form an orthonormal basis, called the *Bell basis*, for a two-qubit space. We can also label the Bell states as

$$|\Phi_{zx}\rangle^{AB} \equiv \left(Z^A\right)^z \left(X^A\right)^x \left|\Phi^+\right\rangle^{AB}, \tag{3.197}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where the two-bit binary number zx indicates whether Alice applies I^A, Z^A, X^A, or ZX^A. Then the states $|\Phi_{00}\rangle^{AB}$, $|\Phi_{01}\rangle^{AB}$, $|\Phi_{10}\rangle^{AB}$, and $|\Phi_{11}\rangle^{AB}$ are in correspondence with the respective states $|\Phi^+\rangle^{AB}$, $|\Psi^+\rangle^{AB}$, $|\Phi^-\rangle^{AB}$, and $|\Psi^-\rangle^{AB}$.

Exercise 3.5.12 Show that the Bell states form an orthonormal basis:

$$\langle \Phi_{zx} | \Phi_{z'x'} \rangle = \delta(z, z')\delta(x, x'). \tag{3.198}$$

Exercise 3.5.13 Show that the following identities hold:

$$|00\rangle^{AB} = \frac{1}{\sqrt{2}}\Big(|\Phi^+\rangle^{AB} + |\Phi^-\rangle^{AB}\Big), \tag{3.199}$$

$$|01\rangle^{AB} = \frac{1}{\sqrt{2}}\Big(|\Psi^+\rangle^{AB} + |\Psi^-\rangle^{AB}\Big), \tag{3.200}$$

$$|10\rangle^{AB} = \frac{1}{\sqrt{2}}\Big(|\Psi^+\rangle^{AB} - |\Psi^-\rangle^{AB}\Big), \tag{3.201}$$

$$|11\rangle^{AB} = \frac{1}{\sqrt{2}}\Big(|\Phi^+\rangle^{AB} - |\Phi^-\rangle^{AB}\Big). \tag{3.202}$$

Exercise 3.5.14 Show that the following identities hold by using the relation in (3.197):

$$|\Phi^+\rangle^{AB} = \frac{1}{\sqrt{2}}\Big(|{+}{+}\rangle^{AB} + |{-}{-}\rangle^{AB}\Big), \tag{3.203}$$

$$|\Phi^-\rangle^{AB} = \frac{1}{\sqrt{2}}\Big(|{-}{+}\rangle^{AB} + |{+}{-}\rangle^{AB}\Big), \tag{3.204}$$

$$|\Psi^+\rangle^{AB} = \frac{1}{\sqrt{2}}\Big(|{+}{+}\rangle^{AB} - |{-}{-}\rangle^{AB}\Big), \tag{3.205}$$

$$|\Psi^-\rangle^{AB} = \frac{1}{\sqrt{2}}\Big(|{-}{+}\rangle^{AB} - |{+}{-}\rangle^{AB}\Big). \tag{3.206}$$

Entanglement is one of the most useful resources in quantum computing, quantum communication, and in the setting of quantum Shannon theory that we explore in this book. Our goal in this book is merely to study entanglement as a resource, but there are many other aspects of entanglement that one can study, such as measures of entanglement, multiparty entanglement, and generalized Bell's inequalities [155].

3.6 Summary and Extensions to Qudit States

We now end our overview of the noiseless quantum theory by summarizing its main postulates in terms of quantum states that are on d-dimensional systems. Such states are called *qudit states*, in analogy with the name "qubit" for two-dimensional quantum systems.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

3.6.1 Qudits

A qudit state $|\psi\rangle$ is an arbitrary superposition of some set of orthonormal basis states $\{|j\rangle\}_{j\in\{0,\ldots,d-1\}}$ for a d-dimensional quantum system:

$$|\psi\rangle \equiv \sum_{j=0}^{d-1} \alpha_j |j\rangle. \tag{3.207}$$

The amplitudes α_j obey the normalization condition $\sum_{j=0}^{d-1} |\alpha_j|^2 = 1$.

3.6.2 Unitary Evolution

The first postulate of the quantum theory is that we can perform a unitary (reversible) evolution U on this state. The resulting state is

$$U|\psi\rangle, \tag{3.208}$$

meaning that we apply the operator U to the state $|\psi\rangle$.

One example of a unitary evolution is the cyclic shift operator $X(x)$ that acts on the orthonormal states $\{|j\rangle\}_{j\in\{0,\ldots,d-1\}}$ as follows:

$$X(x)|j\rangle = |x \oplus j\rangle, \tag{3.209}$$

where $\oplus$ is a cyclic addition operator, meaning that the result of the addition is $(x + j) \bmod(d)$. Notice that the X Pauli operator has a similar behavior on the qubit computational basis states because

$$X|i\rangle = |i \oplus 1\rangle, \tag{3.210}$$

for $i \in \{0, 1\}$. Therefore, the operator $X(x)$ is one qudit analog of the X Pauli operator.

Exercise 3.6.1 Show that the inverse of $X(x)$ is $X(-x)$.

Exercise 3.6.2 Show that the matrix representation $X(x)$ of the $X(x)$ operator is a matrix with elements

$$[X(x)]_{i,j} = \delta_{i,j\oplus x}. \tag{3.211}$$

Another example of a unitary evolution is the *phase operator* $Z(z)$. It applies a state-dependent phase to a basis state. It acts as follows on the qudit computational basis states $\{|j\rangle\}_{j\in\{0,\ldots,d-1\}}$:

$$Z(z)|j\rangle = \exp\{i2\pi zj/d\}|j\rangle. \tag{3.212}$$

This operator is the qudit analog of the Pauli Z operator. The d^2 operators $\{X(x)Z(z)\}_{x,z\in\{0,\ldots,d-1\}}$ are known as the *Heisenberg-Weyl operators*.

Exercise 3.6.3 Show that $Z(1)$ is equivalent to the Pauli Z operator for the case that the dimension $d = 2$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 3.6.4 Show that the inverse of $Z(z)$ is $Z(-z)$.

Exercise 3.6.5 Show that the matrix representation of the phase operator $Z(z)$ is

$$[Z(z)]_{j,k} = \exp\{i2\pi zj/d\}\delta_{j,k}. \tag{3.213}$$

In particular, this result implies that the $Z(z)$ operator has a diagonal matrix representation with respect to the qudit computational basis states $\{|j\rangle\}_{j\in\{0,\ldots,d-1\}}$. Thus, the qudit computational basis states $\{|j\rangle\}_{j\in\{0,\ldots,d-1\}}$ are eigenstates of the phase operator $Z(z)$ (similar to the qubit computational basis states being eigenstates of the Pauli Z operator). The eigenvalue corresponding to the eigenstate $|j\rangle$ is $\exp\{i2\pi zj/d\}$.

Exercise 3.6.6 Show that the eigenstates $|l\rangle_X$ of the cyclic shift operator $X(1)$ are the Fourier-transformed states $|l\rangle_X$ where

$$|l\rangle_X \equiv \frac{1}{\sqrt{d}}\sum_{j=0}^{d-1}\exp\{i2\pi lj/d\}|j\rangle, \tag{3.214}$$

l is an integer in the set $\{0,\ldots,d-1\}$, and the subscript X for the state $|l\rangle_X$ indicates that it is an X eigenstate. Show that the eigenvalue corresponding to the state $|l\rangle_X$ is $\exp\{-i2\pi l/d\}$. Conclude that these states are also eigenstates of the operator $X(x)$, but the corresponding eigenvalues are $\exp\{-i2\pi lx/d\}$.

Exercise 3.6.7 Show that the $+/-$ basis states are a special case of the states in (3.214) when $d = 2$.

Exercise 3.6.8 The Fourier transform operator F is the qudit analog of the Hadamard H. We define it to take Z eigenstates to X eigenstates.

$$F \equiv \sum_{j=0}^{d-1}|j\rangle_X\langle j|_Z, \tag{3.215}$$

where the subscript Z indicates a Z eigenstate. It performs the following transformation on the qudit computational basis states:

$$|j\rangle \rightarrow \frac{1}{\sqrt{d}}\sum_{k=0}^{d-1}\exp\{i2\pi jk/d\}|k\rangle. \tag{3.216}$$

Show that the following relations hold for the Fourier transform operator F:

$$FX(x)F^\dagger = Z(x), \tag{3.217}$$
$$FZ(z)F^\dagger = X(-z). \tag{3.218}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 3.6.9 Show that the commutation relations of the cyclic shift operator $X(x)$ and the phase operator $Z(z)$ are as follows:

$$X(x_1)Z(z_1)X(x_2)Z(z_2) = \exp\{2\pi i(z_1x_2 - x_1z_2)/d\}X(x_2)Z(z_2)X(x_1)Z(z_1). \quad (3.219)$$

You can get this result by first showing that

$$X(x)Z(z) = \exp\{-2\pi izx/d\}Z(z)X(x). \quad (3.220)$$

3.6.3 Measurement of Qudits

Measurement of qudits is similar to measurement of qubits. Suppose that we have some state $|\psi\rangle$. Suppose further that we would like to measure some Hermitian operator A with the following diagonalization:

$$A = \sum_j f(j)\Pi_j, \quad (3.221)$$

where $\Pi_j\Pi_k = \Pi_j\delta_{j,k}$, and $\sum_j \Pi_j = I$. A measurement of the operator A then returns the result j with the following probability:

$$p(j) = \langle\psi|\Pi_j|\psi\rangle, \quad (3.222)$$

and the resulting state is

$$\frac{\Pi_j|\psi\rangle}{\sqrt{p(j)}}. \quad (3.223)$$

The calculation of the expectation of the operator A is similar to how we calculate in the qubit case:

$$\mathbb{E}[A] = \sum_j f(j)\langle\psi|\Pi_j|\psi\rangle \quad (3.224)$$

$$= \langle\psi|\sum_j f(j)\Pi_j|\psi\rangle \quad (3.225)$$

$$= \langle\psi|A|\psi\rangle. \quad (3.226)$$

We give two quick examples of qudit operators that we might like to measure. The operators $X(1)$ and $Z(1)$ are not completely analogous to the respective Pauli X and Pauli Z operators because $X(1)$ and $Z(1)$ are not Hermitian. Thus, we cannot directly measure these operators. Instead, we construct operators that are essentially equivalent to "measuring the operators" $X(1)$ and $Z(1)$. Let us first consider the $Z(1)$ operator. Its eigenstates are the qudit computational basis states $\{|j\rangle\}_{j\in\{0,\ldots,d-1\}}$. We can form the operator $M_{Z(1)}$ as

$$M_{Z(1)} \equiv \sum_{j=0}^{d-1} j|j\rangle\langle j|. \quad (3.227)$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Measuring this operator is equivalent to measuring in the qudit computational basis. The expectation of this operator for a qudit $|\psi\rangle$ in the state in (3.207) is

$$\mathbb{E}\left[M_{Z(1)}\right] = \langle\psi|M_{Z(1)}|\psi\rangle \tag{3.228}$$

$$= \sum_{j'=0}^{d-1}\langle j'|\alpha_{j'}^{*}\sum_{j=0}^{d-1} j|j\rangle\langle j|\sum_{j''=0}^{d-1}\alpha_{j''}|j''\rangle \tag{3.229}$$

$$= \sum_{j',j,j''=0}^{d-1} j\alpha_{j'}^{*}\alpha_{j''}\langle j'|j\rangle\langle j|j''\rangle \tag{3.230}$$

$$= \sum_{j=0}^{d-1} j|\alpha_j|^2. \tag{3.231}$$

Similarly, we can construct an operator $M_{X(1)}$ for "measuring the operator $X(1)$" by using the eigenstates $|j\rangle_X$ of the $X(1)$ operator:

$$M_{X(1)} \equiv \sum_{j=0}^{d-1} j|j\rangle_X\langle j|_X. \tag{3.232}$$

We leave it as an exercise to determine the expectation when measuring the $M_{X(1)}$ operator.

Exercise 3.6.10 Suppose the qudit is in the state $|\psi\rangle$ in (3.207). Show that the expectation of the $M_{X(1)}$ operator is

$$\mathbb{E}\left[M_{X(1)}\right] = \frac{1}{d}\sum_{j=0}^{d-1} j\left|\sum_{j'=0}^{d-1}\alpha_{j'}\exp\{-i2\pi j'j/d\}\right|^2. \tag{3.233}$$

Hint: First show that we can represent the state $|\psi\rangle$ in the $X(1)$ eigenbasis as follows:

$$|\psi\rangle = \sum_{l=0}^{d-1}\frac{1}{\sqrt{d}}\left(\sum_{j=0}^{d-1}\alpha_j\exp\{-i2\pi lj/d\}\right)|l\rangle_X. \tag{3.234}$$

3.6.4 Composite Systems of Qudits

We can define a system of multiple qudits again by employing the tensor product. A general two-qudit state on systems A and B has the following form:

$$|\xi\rangle^{AB} \equiv \sum_{j,k=0}^{d-1}\alpha_{j,k}|j\rangle^A|k\rangle^B. \tag{3.235}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Evolution of two-qudit states is similar as before. Suppose Alice applies a unitary U^A to her qudit. The result is as follows:

$$\left(U^A \otimes I^B\right)|\xi\rangle^{AB} = \left(U^A \otimes I^B\right) \sum_{j,k=0}^{d-1} \alpha_{j,k}|j\rangle^A|k\rangle^B \tag{3.236}$$

$$= \sum_{j,k=0}^{d-1} \alpha_{j,k}\Big(U^A|j\rangle^A\Big)|k\rangle^B, \tag{3.237}$$

which follows by linearity. Bob applying a local unitary U^B has a similar form. The application of some global unitary U^{AB} is as follows:

$$U^{AB}|\xi\rangle^{AB}. \tag{3.238}$$

The Qudit Bell States

Two-qudit states can be entangled as well. The maximally-entangled qudit state is as follows:

$$|\Phi\rangle^{AB} \equiv \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |i\rangle^A|i\rangle^B. \tag{3.239}$$

When Alice possesses the first qudit and Bob possesses the second qudit and they are also separated in space, the above state is a resource known as an *edit* (pronounced "ee · dit"). It is useful in the qudit versions of the teleportation protocol and the super-dense coding protocol discussed in Chapter 6.

Consider applying the operator $X(x)Z(z)$ to Alice's side of the maximally entangled state $|\Phi\rangle^{AB}$. We use the following notation:

$$|\Phi_{x,z}\rangle^{AB} \equiv \left(X^A(x)Z^A(z) \otimes I^B\right)|\Phi\rangle^{AB}. \tag{3.240}$$

The d^2 states $\left\{|\Phi_{x,z}\rangle^{AB}\right\}_{x,z=0}^{d-1}$ are known as the qudit Bell states and are important in qudit quantum protocols and in quantum Shannon theory. Exercise 3.6.11 asks you to verify that these states form a complete, orthonormal basis. Thus, one can measure two qudits in the qudit Bell basis.

Similar to the qubit case, it is straightforward to see that the qudit state can generate a *dit* of common randomness by extending the arguments in Section 3.5.6. We end our review of the noiseless quantum theory with some exercises. The transpose trick is one of our most important tools for manipulating maximally entangled states.

Exercise 3.6.11 Show that the set of states $\left\{|\Phi_{x,z}\rangle^{AB}\right\}_{x,z=0}^{d-1}$ form a complete, orthonormal basis:

$$\langle\Phi_{x',z'}|\Phi_{x,z}\rangle = \delta_{x,x'}\delta_{z,z'}, \tag{3.241}$$

$$\sum_{x,z=0}^{d-1} |\Phi_{x,z}\rangle\langle\Phi_{x,z}| = I^{AB}. \tag{3.242}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 3.6.12 (Transpose Trick) Show that the following "transpose trick" or "ricochet" property holds for a maximally entangled state $|\Phi\rangle^{AB}$ and any matrix M:

$$\left(M^A \otimes I^B\right)|\Phi\rangle^{AB} = \left(I^A \otimes \left(M^T\right)^B\right)|\Phi\rangle^{AB}. \tag{3.243}$$

The implication is that some local action of Alice on $|\Phi\rangle^{AB}$ is equivalent to Bob performing the transpose of this action on his half of the state.

Schmidt decomposition

The Schmidt decomposition is one of the most important tools for analyzing bipartite pure states. The Schmidt decomposition shows that it is possible to decompose any pure, two-qudit state as a superposition of corresponding states. We state this result formally as a theorem.

Theorem 3.6.1 (Schmidt decomposition)**.** *Suppose that we have a two-qudit pure state,*

$$|\psi\rangle^{AB}, \tag{3.244}$$

where the Hilbert spaces for systems A and B have the same dimension d. Then it is possible to express this state as follows:

$$|\psi\rangle^{AB} \equiv \sum_{i=0}^{d-1} \lambda_i |i\rangle^A |i\rangle^B, \tag{3.245}$$

where the amplitudes λ_i are normalized so that $\sum_i |\lambda_i|^2 = 1$, the states $\{|i\rangle^A\}$ form an orthonormal basis for system A, and the states $\{|i\rangle^B\}$ form an orthonormal basis for the system B. The Schmidt rank of a bipartite state is equal to the number of non-zero coefficients λ_i in its Schmidt decomposition.

Proof. We now prove the above theorem. Consider an arbitrary two-qudit pure state $|\psi\rangle^{AB}$. We can express it as follows:

$$|\psi\rangle^{AB} = \sum_{j,k} \alpha_{j,k} |j\rangle^A |k\rangle^B, \tag{3.246}$$

for some amplitudes $\alpha_{j,k}$ and some orthonormal bases $\{|j\rangle^A\}$ and $\{|k\rangle^B\}$ on the respective systems A and B. Let us write the matrix formed by the coefficients $\alpha_{j,k}$ as some matrix A where

$$[A]_{j,k} = \alpha_{j,k}. \tag{3.247}$$

This matrix A admits a singular value decomposition of the form

$$A = U\Lambda V, \tag{3.248}$$

where the matrices U and V are both unitary and the matrix Λ is diagonal. Let us write the matrix elements of U as $u_{j,i}$, those of Λ as λ_i (we only need one index for the elements

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

along the diagonal), and those of V as $v_{i,k}$. The above matrix equation is then equal to the following set of equations:

$$\alpha_{j,k} = \sum_i u_{j,i} \lambda_i v_{i,k}. \tag{3.249}$$

Let us make this substitution into the expression for the state in (3.246):

$$|\psi\rangle^{AB} = \sum_{j,k} \left(\sum_i u_{j,i} \lambda_i v_{i,k} \right) |j\rangle^A |k\rangle^B. \tag{3.250}$$

Readjusting some terms by exploiting the properties of the tensor product, we find that

$$|\psi\rangle^{AB} = \sum_i \lambda_i \left(\sum_j u_{j,i} |j\rangle^A \right) \otimes \left(\sum_k v_{i,k} |k\rangle^B \right) \tag{3.251}$$

$$= \sum_i \lambda_i |i\rangle^A |i\rangle^B, \tag{3.252}$$

where we define the orthonormal basis on the A system as $|i\rangle^A \equiv \sum_j u_{j,i} |j\rangle^A$ and we define the orthonormal basis on the B system as $|i\rangle^B \equiv \sum_k v_{i,k} |k\rangle^B$. This final step completes the proof of the theorem, but the next exercise asks you to verify that the set of states $\{|i\rangle^A\}$ form an orthonormal basis (the proof for the set of states $\{|i\rangle^B\}$ is similar). □

Remark 3.6.1 The Schmidt decomposition applies not only to bipartite systems but to any number of systems where we can make a bipartite cut of the systems. For example, suppose that there is a state $|\phi\rangle^{ABCDE}$ on systems $ABCDE$. We could say that AB are part of one system and CDE are part of another system and write a Schmidt decomposition for this state as follows:

$$|\phi\rangle^{ABCDE} = \sum_y \sqrt{p_Y(y)} |y\rangle^{AB} |y\rangle^{CDE}, \tag{3.253}$$

where $\{|y\rangle^{AB}\}$ is an orthonormal basis for the joint system AB and $\{|y\rangle^{CDE}\}$ is an orthonormal basis for the joint system CDE.

Exercise 3.6.13 Verify that the set of states $\{|i\rangle^A\}$ form an orthonormal basis by exploiting the unitarity of the matrix U.

3.7 History and Further Reading

There are many great books on quantum mechanics that outline the mathematical background. The books of Bohm [41], Sakurai [209], and Nielsen and Chuang [197] are among these. The ideas for the resource inequality formalism first appeared in a popular article of Bennett [19] and another of his papers [20]. The no-deletion theorem is in Ref. [202]. The review article of the Horodecki family is a useful reference on the study of entanglement [155].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 4

The Noisy Quantum Theory

In general, we may not know for certain whether we possess a particular quantum state. Instead, we may only have a probabilistic description of an ensemble of quantum states. This chapter re-establishes the postulates of the quantum theory so that they incorporate a lack of complete information about a quantum system. The density operator formalism is a powerful mathematical tool for describing this scenario. This chapter also establishes how to model the noisy evolution of a quantum system, and we explore models of noisy quantum channels that are the analogs of the noisy classical channel discussed in Section 2.2.3 of Chapter 2.

You might have noticed that the development in the previous chapter relied on the premise that the possessor of a quantum system has perfect knowledge of the state of a given system. For instance, we assumed that Alice knows that she possesses a qubit in the state $|\psi\rangle$ where

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle. \tag{4.1}$$

Also, we assumed that Alice and Bob might know that they share an ebit $|\Phi^+\rangle$. We even assumed perfect knowledge of a unitary evolution or a particular measurement that a possessor of a quantum state may apply to it.

This assumption of perfect, definite knowledge of a quantum state is a difficult one to justify in practice. In reality, it is difficult to prepare, evolve, or measure a quantum state exactly as we wish. Slight errors may occur in the preparation, evolution, or measurement due to imprecise devices or to coupling with other degrees of freedom outside of the system that we are controlling. An example of such imprecision can occur in the coupling of two photons at a beamsplitter. We may not be able to tune the reflectivity of the beamsplitter exactly or may not have the timing of the arrival of the photons exactly set. The noiseless quantum theory as we presented it in the previous section cannot handle such imprecisions.

In this chapter, we relax the assumption of perfect knowledge of the preparation, evolution, or measurement of quantum states and develop a noisy quantum theory that incorporates an imprecise knowledge of these states. The noisy quantum theory fuses probability theory and the quantum theory into one formalism.

We proceed with the development of the noisy quantum theory in the following order:

1. We first present the density operator formalism, which gives a representation for a noisy, imprecise quantum state.

2. We then discuss the general form of measurements and the effect of them on our description of a noisy quantum state. We specifically discuss the POVM (positive operator-valued measure) formalism that gives a more general way of describing measurements.

3. We proceed to composite noisy systems, which admit a particular form, and we discuss several possible states of composite noisy systems including product states, separable states, classical-quantum states, entangled states, and arbitrary states.

4. Next, we consider the Kraus representation of a noisy quantum channel, which gives a way to describe noisy evolution, and we discuss important examples of noisy quantum channels.

4.1 Noisy Quantum States

We generally may not have perfect knowledge of a prepared quantum state. Suppose a third party, Bob, prepares a state for us and only gives us a probabilistic description of it. We may only know that Bob selects the state $|\psi_x\rangle$ with a certain probability $p_X(x)$. Our description of the state is then with an ensemble $\mathcal{E}$ of quantum states where

$$\mathcal{E} \equiv \{p_X(x), |\psi_x\rangle\}_{x\in\mathcal{X}}. \tag{4.2}$$

In the above, X is a random variable with distribution $p_X(x)$. Each realization x of random variable X belongs to an alphabet $\mathcal{X}$. For our purposes, it is sufficient for us to say that $\mathcal{X} \equiv \{1, \ldots, |\mathcal{X}|\}$. Thus, the realization x merely acts as an index, meaning that the quantum state is $|\psi_x\rangle$ with probability $p_X(x)$. We also assume that each state $|\psi_x\rangle$ is a qudit state that lives on a system of dimension d.

A simple example is the following ensemble:

$$\left\{\left\{\frac{1}{3}, |1\rangle\right\}, \left\{\frac{2}{3}, |3\rangle\right\}\right\}. \tag{4.3}$$

The states $|1\rangle$ and $|3\rangle$ live on a four-dimensional system with basis states

$$\{|0\rangle, |1\rangle, |2\rangle, |3\rangle\}. \tag{4.4}$$

The interpretation of this ensemble is that the state is $|1\rangle$ with probability $1/3$ and the state is $|3\rangle$ with probability $2/3$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

4.1.1 The Density Operator

Suppose now that we have the ability to perform a perfect measurement of a system with ensemble description $\mathcal{E}$ in (4.2). Let Π_j be the elements of this projective measurement so that $\sum_j \Pi_j = I$, and let J be the random variable that denotes the index j of the measurement outcome. Let us suppose at first, without loss of generality, that the state in the ensemble is $|\psi_x\rangle$ for some $x \in \mathcal{X}$. Then the Born rule of the noiseless quantum theory states that the conditional probability $p_{J|X}(j|x)$ of obtaining measurement result j (given that the state is $|\psi_x\rangle$) is

$$p_{J|X}(j|x) = \langle\psi_x|\Pi_j|\psi_x\rangle, \tag{4.5}$$

and the post-measurement state is

$$\frac{\Pi_j|\psi_x\rangle}{\sqrt{p_{J|X}(j|x)}}. \tag{4.6}$$

But, we would also like to know the actual probability $p_J(j)$ of obtaining measurement result j for the ensemble description $\mathcal{E}$. By the *law of total probability*, the unconditional probability $p_J(j)$ is

$$p_J(j) = \sum_{x\in\mathcal{X}} p_{J|X}(j|x)p_X(x) \tag{4.7}$$

$$= \sum_{x\in\mathcal{X}} \langle\psi_x|\Pi_j|\psi_x\rangle p_X(x). \tag{4.8}$$

The *trace* $\text{Tr}\{A\}$ of an operator A is

$$\text{Tr}\{A\} \equiv \sum_i \langle i|A|i\rangle, \tag{4.9}$$

where $|i\rangle$ is some complete, orthonormal basis. (Observe that the trace operation is *linear*.) We can then show the following useful property with the above definition:

$$\text{Tr}\{\Pi_j|\psi_x\rangle\langle\psi_x|\} = \sum_i \langle i|\Pi_j|\psi_x\rangle\langle\psi_x|i\rangle \tag{4.10}$$

$$= \sum_i \langle\psi_x|i\rangle\langle i|\Pi_j|\psi_x\rangle \tag{4.11}$$

$$= \langle\psi_x|\left(\sum_i |i\rangle\langle i|\right)\Pi_j|\psi_x\rangle \tag{4.12}$$

$$= \langle\psi_x|\Pi_j|\psi_x\rangle. \tag{4.13}$$

The last equality uses the completeness relation $\sum_i |i\rangle\langle i| = I$. Thus, we continue with the development in (4.8) and show that

$$p_J(j) = \sum_{x\in\mathcal{X}} \text{Tr}\{\Pi_j|\psi_x\rangle\langle\psi_x|\}p_X(x) \tag{4.14}$$

$$= \text{Tr}\left\{\Pi_j \sum_{x\in\mathcal{X}} p_X(x)|\psi_x\rangle\langle\psi_x|\right\}. \tag{4.15}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can rewrite the last equation as follows:

$$p_J(j) = \text{Tr}\{\Pi_j \rho\}, \tag{4.16}$$

where we define the *density operator* ρ as

$$\rho \equiv \sum_{x \in \mathcal{X}} p_X(x) |\psi_x\rangle\langle\psi_x|. \tag{4.17}$$

The above operator is known as the density operator because it is the quantum analog of a probability density function.

We sometimes refer to the density operator as the *expected density operator* because there is a sense in which we are taking the expectation over all of the states in the ensemble in order to obtain the density operator. We can equivalenty write the density operator as follows:

$$\rho = \mathbb{E}_X\{|\psi_X\rangle\langle\psi_X|\}, \tag{4.18}$$

where the expectation is with respect to the random variable X. Note that we are careful to use the notation $|\psi_X\rangle$ instead of the notation $|\psi_x\rangle$ for the state inside of the expectation because the state $|\psi_X\rangle$ is a random quantum state, random with respect to a classical random variable X.

Exercise 4.1.1 Suppose the ensemble has a degenerate probability distribution, say $p_X(0) = 1$ and $p_X(x) = 0$ for all $x \neq 0$. What is the density operator of this degenerate ensemble?

Properties of the Density Operator

What are the properties that a given density operator must satisfy? Let us consider taking the trace of ρ:

$$\begin{aligned}
\text{Tr}\{\rho\} &= \text{Tr}\left\{\sum_{x \in \mathcal{X}} p_X(x) |\psi_x\rangle\langle\psi_x|\right\} && (4.19)\\
&= \sum_{x \in \mathcal{X}} p_X(x) \text{Tr}\{|\psi_x\rangle\langle\psi_x|\} && (4.20)\\
&= \sum_{x \in \mathcal{X}} p_X(x) \langle\psi_x|\psi_x\rangle && (4.21)\\
&= \sum_{x \in \mathcal{X}} p_X(x) && (4.22)\\
&= 1. && (4.23)
\end{aligned}$$

The above development shows that every density operator should have *unit trace* because it arises from an ensemble of quantum states. Every density operator is also *positive*, meaning that

$$\forall |\varphi\rangle : \qquad \langle\varphi|\rho|\varphi\rangle \geq 0. \tag{4.24}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We write $\rho \geq 0$ to indicate that an operator is positive. The proof of positivity of any density operator ρ is as follows:

$$\langle\varphi|\rho|\varphi\rangle = \langle\varphi|\left(\sum_{x\in\mathcal{X}} p_X(x)|\psi_x\rangle\langle\psi_x|\right)|\varphi\rangle \tag{4.25}$$

$$= \sum_{x\in\mathcal{X}} p_X(x)\langle\varphi|\psi_x\rangle\langle\psi_x|\varphi\rangle \tag{4.26}$$

$$= \sum_{x\in\mathcal{X}} p_X(x)|\langle\varphi|\psi_x\rangle|^2 \geq 0. \tag{4.27}$$

The inequality follows because each $p_X(x)$ is a probability and is therefore non-negative.

Let us consider taking the conjugate transpose of the density operator ρ:

$$\rho^\dagger = \left(\sum_{x\in\mathcal{X}} p_X(x)|\psi_x\rangle\langle\psi_x|\right)^\dagger \tag{4.28}$$

$$= \sum_{x\in\mathcal{X}} p_X(x)(|\psi_x\rangle\langle\psi_x|)^\dagger \tag{4.29}$$

$$= \sum_{x\in\mathcal{X}} p_X(x)|\psi_x\rangle\langle\psi_x| \tag{4.30}$$

$$= \rho. \tag{4.31}$$

Every density operator is thus a *Hermitian* operator as well because the conjugate transpose of ρ is ρ.

Ensembles and the Density Operator

Every ensemble has a unique density operator, but the opposite does not necessarily hold: every density operator does not correspond to a unique ensemble and could correspond to many ensembles.

Exercise 4.1.2 Show that the ensembles

$$\{\{1/2, |0\rangle\}, \{1/2, |1\rangle\}\} \tag{4.32}$$

and

$$\{\{1/2, |+\rangle\}, \{1/2, |-\rangle\}\} \tag{4.33}$$

have the same density operator.

This last result has profound implications for the predictions of the quantum theory because it is possible for two or more completely different ensembles to have the same probabilities for measurement results. It also has important implications for quantum Shannon theory as well.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

By the spectral theorem, it follows that every density operator ρ has a spectral decomposition in terms of its eigenstates $\{|\phi_x\rangle\}_{x\in\{0,\ldots,d-1\}}$ because every ρ is Hermitian:

$$\rho = \sum_{x=0}^{d-1} \lambda_x |\phi_x\rangle\langle\phi_x|, \tag{4.34}$$

where the coefficients λ_x are the eigenvalues.

Exercise 4.1.3 Show that the coefficients λ_x are probabilities using the facts that $\text{Tr}\{\rho\} = 1$ and $\rho \geq 0$.

Thus, given any density operator ρ, we can define a "canonical" ensemble $\{\lambda_x, |\phi_x\rangle\}$ corresponding to it. This observation is so important for quantum Shannon theory that we see this idea arise again and again throughout this book. The ensemble arising from the spectral theorem is the most "efficient" ensemble, in a sense, and we will explore this idea more in Chapter 17 on quantum compression (known as Schumacher compression after its inventor).

Density Operator as the State

We can also refer to the density operator as the *state* of a given quantum system because it is possible to use it to calculate all of the predictions of the quantum theory. We can make these calculations without having an ensemble description—all we need is the density operator. The noisy quantum theory also subsumes the noiseless quantum theory because any state $|\psi\rangle$ has a corresponding density operator $|\psi\rangle\langle\psi|$ in the noisy quantum theory, and all calculations with this density operator in the noisy quantum theory give the same results as using the state $|\psi\rangle$ in the noiseless quantum theory. For these reasons, we will say that the *state* of a given quantum system is a density operator.

One of the most important states in the noisy quantum theory is the maximally mixed state π. The maximally mixed state π arises as the density operator of a uniform ensemble of orthogonal states $\left\{\frac{1}{d}, |x\rangle\right\}$, where d is the dimensionality of the Hilbert space. The maximally mixed state π is then equal to

$$\pi = \frac{1}{d}\sum_{x\in\mathcal{X}} |x\rangle\langle x| = \frac{I}{d}. \tag{4.35}$$

Exercise 4.1.4 Show that π is the density operator of the ensemble that chooses $|0\rangle$, $|1\rangle$, $|+\rangle$, $|-\rangle$ with equal probability.

The purity $P(\rho)$ of a density operator ρ is equal to

$$P(\rho) \equiv \text{Tr}\{\rho^\dagger \rho\} = \text{Tr}\{\rho^2\}. \tag{4.36}$$

The purity is one particular measure of the noisiness of a quantum state. The purity of a pure state is equal to one, and the purity of a mixed state is strictly less than one.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The Density Operator on the Bloch Sphere

Consider that the density operator of a pure qubit state

$$|\psi\rangle \equiv \cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\varphi}\sin\left(\frac{\theta}{2}\right)|1\rangle \tag{4.37}$$

has the following density operator representation:

$$\begin{aligned}|\psi\rangle\langle\psi| &= \left(\cos\left(\frac{\theta}{2}\right)|0\rangle + e^{i\varphi}\sin\left(\frac{\theta}{2}\right)|1\rangle\right)\left(\cos\left(\frac{\theta}{2}\right)\langle 0| + e^{-i\varphi}\sin\left(\frac{\theta}{2}\right)\langle 1|\right) &(4.38)\\ &= \cos^2\left(\frac{\theta}{2}\right)|0\rangle\langle 0| + e^{-i\varphi}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)|0\rangle\langle 1| \\ &\quad + e^{i\varphi}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right)|1\rangle\langle 0| + \sin^2\left(\frac{\theta}{2}\right)|1\rangle\langle 1|. &(4.39)\end{aligned}$$

The matrix representation, or *density matrix*, of this density operator with respect to the computational basis is as follows:

$$\begin{bmatrix} \cos^2\left(\frac{\theta}{2}\right) & e^{-i\varphi}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right) \\ e^{i\varphi}\sin\left(\frac{\theta}{2}\right)\cos\left(\frac{\theta}{2}\right) & \sin^2\left(\frac{\theta}{2}\right)\end{bmatrix}. \tag{4.40}$$

Using trigonometric identities, it follows that the density matrix is equal to the following matrix:

$$\frac{1}{2}\begin{bmatrix} 1+\cos(\theta) & \sin(\theta)(\cos(\varphi) - i\sin(\varphi)) \\ \sin(\theta)(\cos(\varphi) + i\sin(\varphi)) & 1-\cos(\theta)\end{bmatrix}. \tag{4.41}$$

We can further exploit the Pauli matrices, defined in Section 3.3.3, to represent the density matrix as follows:

$$\frac{1}{2}(I + r_x X + r_y Y + r_z Z), \tag{4.42}$$

where

$$\begin{aligned} r_x &= \sin(\theta)\cos(\varphi), &(4.43)\\ r_y &= \sin(\theta)\sin(\varphi), &(4.44)\\ r_z &= \cos(\theta). &(4.45)\end{aligned}$$

The coefficients r_x, r_y, and r_z are none other than the Cartesian representation of the angles θ and φ, and they thus correspond to a unit vector.

More generally, the formula in (4.42) can represent an arbitrary density operator where the coefficients r_x, r_y, and r_z do not necessarily correspond to a unit vector, but rather a vector $\mathbf{r}$ such that $\|\mathbf{r}\|_2 \leq 1$. Consider that the density matrix in (4.42) is as follows:

$$\frac{1}{2}\begin{bmatrix} 1+r_z & r_x - ir_y \\ r_x + ir_y & 1-r_z\end{bmatrix}. \tag{4.46}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above matrix corresponds to a valid density matrix because it has unit trace, it is Hermitian, and it is positive (the next exercise asks you to verify these facts). This alternate representation of the density matrix as a vector in the Bloch sphere is useful for visualizing noisy processes in the noisy quantum theory.

Exercise 4.1.5 Show that the matrix in (4.46) has unit trace, is Hermitian, and is positive for all $\mathbf{r}$ such that $\|\mathbf{r}\|_2 \leq 1$. It thus corresponds to any valid density matrix.

Exercise 4.1.6 Show that we can compute the Bloch sphere coordinates r_x, r_y, and r_z with the respective formulas $\text{Tr}\{\rho X\}$, $\text{Tr}\{\rho Y\}$, and $\text{Tr}\{\rho Z\}$ using the representation in (4.46) and the result of Exercise 3.3.5.

Exercise 4.1.7 Show that the eigenvalues of a general qubit density operator with density matrix representation in (4.46) are as follows:

$$\frac{1}{2}(1 \pm \|\mathbf{r}\|_2). \tag{4.47}$$

Exercise 4.1.8 Show that a mixture of pure states $|\psi_j\rangle$ each with Bloch vector $\mathbf{r}_j$ and probability $p(j)$ gives a density matrix with the Bloch vector $\mathbf{r}$ where

$$\mathbf{r} = \sum_j p(j)\mathbf{r}_j. \tag{4.48}$$

4.1.2 An Ensemble of Ensembles

The most general ensemble that we can construct is an *ensemble of ensembles*, i.e., an ensemble $\mathcal{F}$ of density operators where

$$\mathcal{F} \equiv \{p_X(x), \rho_x\}. \tag{4.49}$$

The ensemble $\mathcal{F}$ essentially has two layers of randomization. The first layer is from the distribution $p_X(x)$. Each density operator ρ_x in $\mathcal{F}$ arises from an ensemble $\{p_{Y|X}(y|x), |\psi_{xy}\rangle\}$. The conditional distribution $p_{Y|X}(y|x)$ represents the second layer of randomization. Each ρ_x is a density operator with respect to the above ensemble:

$$\rho_x \equiv \sum_y p_{Y|X}(y|x)|\psi_{xy}\rangle\langle\psi_{xy}|. \tag{4.50}$$

The ensemble $\mathcal{F}$ has its own density operator ρ where

$$\rho \equiv \sum_{x,y} p_{Y|X}(y|x)p_X(x)|\psi_{xy}\rangle\langle\psi_{xy}| \tag{4.51}$$

$$= \sum_x p_X(x)\rho_x. \tag{4.52}$$

The density operator ρ is the density operator from the perspective of someone who is ignorant of x. Figure 4.1 displays the process by which we can select the ensemble $\mathcal{F}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

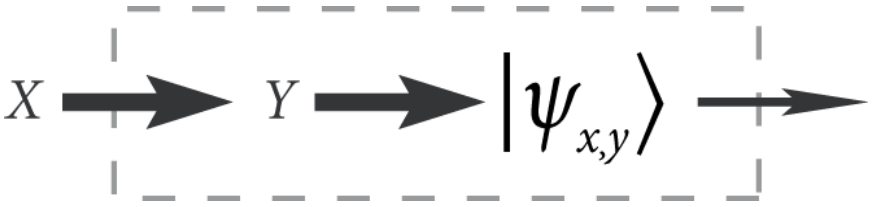

Figure 4.1: The mixing process by which we can generate an "ensemble of ensembles." First choose a realization x according to distribution $p_X(x)$. Then choose a realization y according to the conditional distribution $p_{Y|X}(y|x)$. Finally, choose a state $|\psi_{x,y}\rangle$ according to the realizations x and y. This leads to an ensemble $\{p_X(x), \rho_x\}$ where $\rho_x \equiv \sum_y p_{Y|X}(y|x)|\psi_{x,y}\rangle\langle\psi_{x,y}|$.

4.1.3 Noiseless Evolution of an Ensemble

Quantum states can evolve in a noiseless fashion either according to a unitary operator or a measurement. In this section, we determine the noiseless evolution of an ensemble and its corresponding density operator. (We consider noisy evolution in Section 4.4.)

Noiseless Unitary Evolution of a Noisy State

We first consider noiseless evolution according to some unitary U. Suppose we have the ensemble $\mathcal{E}$ in (4.2) with density operator ρ. Suppose without loss of generality that the state is $|\psi_x\rangle$. Then the evolution postulate of the noiseless quantum theory gives that the state after the unitary evolution is as follows:

$$U|\psi_x\rangle. \tag{4.53}$$

This result implies that the evolution leads to a new ensemble

$$\mathcal{E}_U \equiv \{p_X(x), U|\psi_x\rangle\}_{x\in\mathcal{X}}. \tag{4.54}$$

The density operator of the evolved ensemble is

$$\sum_{x\in\mathcal{X}} p_X(x) U|\psi_x\rangle\langle\psi_x|U^\dagger = U\left(\sum_{x\in\mathcal{X}} p_X(x)|\psi_x\rangle\langle\psi_x|\right)U^\dagger \tag{4.55}$$

$$= U\rho U^\dagger. \tag{4.56}$$

Thus, the above relation shows that we can keep track of the evolution of the density operator ρ, rather than worrying about keeping track of the evolution of every state in the ensemble $\mathcal{E}$. It suffices to keep track of only the density operator evolution because this operator is sufficient to determine the predictions of the quantum theory.

Noiseless Measurement of a Noisy State

In a similar fashion, we can analyze the result of a measurement on a system with ensemble description $\mathcal{E}$ in (4.2). Suppose that we perform a projective measurement with projection

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

operators $\{\Pi_j\}_j$ where $\sum_j \Pi_j = I$. Suppose further without loss of generality that the state in the ensemble is $|\psi_x\rangle$. Then the noiseless quantum theory predicts that the probability of obtaining outcome j conditioned on the index x is

$$p_{J|X}(j|x) = \langle\psi_x|\Pi_j|\psi_x\rangle, \tag{4.57}$$

and the resulting state is

$$\frac{\Pi_j|\psi_x\rangle}{\sqrt{p_{J|X}(j|x)}}. \tag{4.58}$$

Supposing that we receive outcome j, then we have a new ensemble:

$$\mathcal{E}_j \equiv \left\{p_{X|J}(x|j), \Pi_j|\psi_x\rangle/\sqrt{p_{J|X}(j|x)}\right\}_{x\in\mathcal{X}}. \tag{4.59}$$

The density operator for this ensemble is

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} p_{X|J}(x|j)\frac{\Pi_j|\psi_x\rangle\langle\psi_x|\Pi_j}{p_{J|X}(j|x)} \\
&= \Pi_j\left(\sum_{x\in\mathcal{X}}\frac{p_{X|J}(x|j)}{p_{J|X}(j|x)}|\psi_x\rangle\langle\psi_x|\right)\Pi_j &(4.60)\\
&= \Pi_j\left(\sum_{x\in\mathcal{X}}\frac{p_{J|X}(j|x)p_X(x)}{p_{J|X}(j|x)p_J(j)}|\psi_x\rangle\langle\psi_x|\right)\Pi_j &(4.61)\\
&= \frac{\Pi_j\left(\sum_{x\in\mathcal{X}} p_X(x)|\psi_x\rangle\langle\psi_x|\right)\Pi_j}{p_J(j)} &(4.62)\\
&= \frac{\Pi_j\rho\Pi_j}{p_J(j)}. &(4.63)
\end{aligned}$$

The second equality follows from applying the Bayes rule:

$$p_{X|J}(x|j) = p_{J|X}(j|x)p_X(x)/p_J(j). \tag{4.64}$$

The above expression gives the evolution of the density operator under a measurement. We can again employ the law of total probability to compute that $p_J(j)$ is

$$\begin{aligned}
p_J(j) &= \sum_{x\in\mathcal{X}} p_{J|X}(j|x)p_X(x) &(4.65)\\
&= \sum_{x\in\mathcal{X}} p_X(x)\langle\psi_x|\Pi_j|\psi_x\rangle &(4.66)\\
&= \sum_{x\in\mathcal{X}} p_X(x)\mathrm{Tr}\{|\psi_x\rangle\langle\psi_x|\Pi_j\} &(4.67)\\
&= \mathrm{Tr}\left\{\sum_{x\in\mathcal{X}} p_X(x)|\psi_x\rangle\langle\psi_x|\Pi_j\right\} &(4.68)\\
&= \mathrm{Tr}\{\rho\Pi_j\}. &(4.69)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can think of $\text{Tr}\{\rho\Pi_j\}$ as the area of the shadow of ρ onto the space that the projector Π_j projects.

4.1.4 Probability Theory as a Special Case of the Noisy Quantum Theory

It may help to build some intuition for the noisy quantum theory by showing how it contains probability theory as a special case. Indeed, we should expect this containment of probability theory within the noisy quantum theory to hold if the noisy quantum theory is making probabilistic predictions about the physical world.

Let us again begin with an ensemble of quantum states, but this time, let us pick the states in the ensemble to be special states, where they are all orthogonal to one another. If the states in the ensemble are all orthogonal to one another, then they are essentially classical states because there is a measurement that distinguishes them from one another. So, let us pick the ensemble to be $\{p_X(x), |x\rangle\}_{x\in\mathcal{X}}$ where the states $\{|x\rangle\}_{x\in\mathcal{X}}$ form an orthonormal basis for a Hilbert space of dimension $|\mathcal{X}|$. These states are classical because a measurement with the following projection operators can distinguish them:

$$\{|x\rangle\langle x|\}_{x\in\mathcal{X}}. \tag{4.70}$$

The formal analogy of a probability distribution in the quantum world is the density operator:

$$p_X(x) \leftrightarrow \rho. \tag{4.71}$$

The reason for this analogy is that we can use the density operator to calculate expectations and moments of observables.

The formal analogy of a random variable is an observable. Let us consider the following observable:

$$X \equiv \sum_{x\in\mathcal{X}} x|x\rangle\langle x|, \tag{4.72}$$

analogous to the observable in (3.227). We perform the following calculation to determine the expectation of the observable X:

$$\mathbb{E}[X] = \text{Tr}\{X\rho\}. \tag{4.73}$$

Explicitly calculating this quantity, we discover that it is the same computation as that for the expectation of random variable X:

$$\text{Tr}\{X\rho\} = \text{Tr}\left\{\sum_{x\in\mathcal{X}} x|x\rangle\langle x| \sum_{x'\in\mathcal{X}} p_X(x')|x'\rangle\langle x'|\right\} \tag{4.74}$$

$$= \sum_{x,x'\in\mathcal{X}} x\ p_X(x')|\langle x|x'\rangle|^2 \tag{4.75}$$

$$= \sum_{x\in\mathcal{X}} x\ p_X(x). \tag{4.76}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Another useful notion in probability theory is the notion of an indicator random variable $I_A(X)$. We define the indicator function $I_A(x)$ as follows:

$$I_A(x) \equiv \begin{cases} 1 & : \quad x \in A \\ 0 & : \quad x \notin A \end{cases}. \tag{4.77}$$

The expectation $\mathbb{E}[I_A(X)]$ of the indicator random variable $I_A(X)$ is

$$\mathbb{E}[I_A(X)] = \sum_{x \in A} p_X(x). \tag{4.78}$$

$$= p_X(A), \tag{4.79}$$

where $p_X(A)$ represents the probability of the set A. In the quantum theory, we can formulate an indicator observable $I_A(X)$:

$$I_A(X) \equiv \sum_{x \in A} |x\rangle\langle x|. \tag{4.80}$$

It has eigenvalues equal to unity for all eigenvectors with labels x in the set A, and it has null eigenvalues for those eigenvectors with labels outside of A. It is straightforward to show that the expectation $\text{Tr}\{\rho I_A(X)\}$ of the indicator observable $I_A(X)$ is $p_X(A)$.

You may have noticed that the indicator observable is also a projection operator. So, according to the postulates of the quantum theory, we can perform a measurement with elements:

$$\{I_A(X), I_{A^c}(X) \equiv I - I_A(X)\}. \tag{4.81}$$

The result of such a projective measurement is to project onto the subspace given by $I_A(X)$ with probability $p_X(A)$ and to project onto the complementary subspace given by $I_{A^c}(X)$ with probability $1 - p_X(A)$.

We highlight the connection between the noisy quantum theory and probability theory with two more examples. First, suppose that we have two *disjoint* sets A and B. Then the probability of their union is the sum of the probabilities of the individual sets:

$$p(A \cup B) = p(A) + p(B), \tag{4.82}$$

and the probability of the complementary set $(A \cup B)^c = A^c \cap B^c$ is equal to $1 - p(A) - p(B)$. We can perform the analogous calculation in the noisy quantum theory. Let us consider two projection operators

$$\Pi_A \equiv \sum_{x \in A} |x\rangle\langle x|, \tag{4.83}$$

$$\Pi_B \equiv \sum_{x \in B} |x\rangle\langle x|. \tag{4.84}$$

The sum of these projection operators gives a projection onto the union set $A \cup B$:

$$\Pi_{A \cup B} \equiv \sum_{x \in A \cup B} |x\rangle\langle x| = \Pi_A + \Pi_B. \tag{4.85}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 4.1.9 Show that $\text{Tr}\{\Pi_{A\cup B}\rho\} = p(A) + p(B)$ whenever the projectors Π_A and Π_B do not have overlapping support and the density operator ρ is diagonal in the same basis as Π_A and Π_B.

We can also consider intersections of sets. Suppose that we have two sets A and B. The intersection of these two sets consists of all the elements that are common to both sets. There is an associated probability $p(A \cap B)$ with the intersection. We can again formulate this idea in the noisy quantum theory. Consider the projection operators in (4.83-4.84). The multiplication of these two projectors gives a projector onto the intersection of the two spaces:

$$\Pi_{A\cap B} = \Pi_A \Pi_B. \tag{4.86}$$

Exercise 4.1.10 Show that $\text{Tr}\{\Pi_A \Pi_B \rho\} = p(A \cap B)$ whenever the density operator ρ is diagonal in the same basis as Π_A and Π_B.

Such ideas and connections to the classical world are crucial for understanding quantum Shannon theory. Many times, we will be thinking about unions of disjoint subspaces and it is useful to make the analogy with a union of disjoint sets. Also, in Chapter 16 on the Covering Lemma, we will use projection operators to remove some of the support of an operator, and this operation is analogous to taking intersections of sets.

Despite the fact that there is a strong connection for classical states, some of this intuition breaks down by considering the non-orthogonality of quantum states. For example, consider the case of the projectors $\Pi_0 \equiv |0\rangle\langle 0|$ and $\Pi_+ \equiv |+\rangle\langle +|$. The two subspaces onto which these operators project do not intersect, yet we know that the projectors have some overlap because their corresponding states are non-orthogonal. One analogy of the intersection operation is to slice out the support of one operator from another. For example, we can form the operator

$$\Pi_0 \Pi_+ \Pi_0, \tag{4.87}$$

and placing Π_0 on the outside slices out the support of Π_+ that is not in Π_0. Similarly, we can slice out the support of Π_0 not in Π_+ by forming the following operator

$$\Pi_+ \Pi_0 \Pi_+. \tag{4.88}$$

If the two projectors were to commute, then this ordering would not matter, and the resulting operator would be a projector onto the intersection of the two subspaces. But this is not the case for our example here, and the resulting operators are quite different.

Exercise 4.1.11 (Union Bound) Prove a union bound for commuting projectors Π_1 and Π_2 where $0 \leq \Pi_1, \Pi_2 \leq I$ and for an *arbitrary* density operator ρ (not necessarily diagonal in the same basis as Π_1 and Π_2):

$$\text{Tr}\{(I - \Pi_1\Pi_2)\rho\} \leq \text{Tr}\{(I - \Pi_1)\rho\} + \text{Tr}\{(I - \Pi_2)\rho\}. \tag{4.89}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

4.2 Measurement in the Noisy Quantum Theory

We have described measurement in the quantum theory using a set of projectors that form a resolution of the identity. For example, the set $\{\Pi_j\}_j$ of projectors that satisfy the condition $\sum_j \Pi_j = I$ form a valid von Neumann quantum measurement. A projective measurement is not the most general measurement that we can perform on a quantum system (though it is certainly one valid type of quantum measurement).

The most general quantum measurement consists of a set of measurement operators $\{M_j\}_j$ that satisfy the following completeness condition:

$$\sum_j M_j^\dagger M_j = I. \tag{4.90}$$

Suppose that we have a pure state $|\psi\rangle$. Given a set of measurement operators of the above form, the probability for obtaining outcome j is

$$p(j) \equiv \langle\psi|M_j^\dagger M_j|\psi\rangle, \tag{4.91}$$

and the post-measurement state when we receive outcome j is

$$\frac{M_j|\psi\rangle}{\sqrt{p(j)}}. \tag{4.92}$$

Suppose that we instead have an ensemble $\{p_X(x), |\psi_x\rangle\}$ with density operator ρ. We can carry out the same analysis in (4.63) to show that the post-measurement state when we measure result j is

$$\frac{M_j \rho M_j^\dagger}{p(j)} \tag{4.93}$$

where the probability $p(j)$ for obtaining outcome j is

$$p(j) \equiv \text{Tr}\Big\{ M_j^\dagger M_j \rho \Big\}. \tag{4.94}$$

4.2.1 POVM Formalism

Sometimes, we simply may not care about the post-measurement state of a quantum measurement, but instead we only care about the probability for obtaining a particular outcome. This situation arises in the transmission of classical data over a quantum channel. In this situation, we are merely concerned with minimizing the error probabilities of the classical transmission. The receiver does not care about the post-measurement state because he no longer needs it in the quantum information processing protocol.

We can specify a measurement of this sort by some set $\{\Lambda_j\}_j$ of operators that satisfy positivity and completeness:

$$\Lambda_j \geq 0, \tag{4.95}$$

$$\sum_j \Lambda_j = I. \tag{4.96}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The set $\{\Lambda_j\}_j$ of operators is a positive operator-valued measure (POVM). The probability for obtaining outcome j is

$$\langle\psi|\Lambda_j|\psi\rangle, \tag{4.97}$$

if the state is some pure state $|\psi\rangle$. The probability for obtaining outcome j is

$$\text{Tr}\{\Lambda_j\rho\}, \tag{4.98}$$

if the state is in a mixed state described by some density operator ρ.

Exercise 4.2.1 Consider the following five "Chrysler" states:

$$|e_k\rangle \equiv \cos\left(\frac{2\pi k}{5}\right)|0\rangle + \sin\left(\frac{2\pi k}{5}\right)|1\rangle, \tag{4.99}$$

where $k \in \{0, \ldots, 4\}$. These states are the "Chrysler" states because they form a pentagon on the XZ-plane of the Bloch sphere. Show that the following set of states form a valid POVM:

$$\left\{\frac{2}{5}|e_k\rangle\langle e_k|\right\}. \tag{4.100}$$

Exercise 4.2.2 Suppose we have an ensemble $\{p(x), \rho_x\}$ of density operators and a POVM with elements $\{\Lambda_x\}$ that should identify the states ρ_x with high probability, i.e., we would like $\text{Tr}\{\rho_x\Lambda_x\}$ to be as high as possible. The expected success probability of the POVM is then

$$\sum_x p(x)\text{Tr}\{\rho_x\Lambda_x\}. \tag{4.101}$$

Suppose that there exists some operator τ such that

$$\tau \geq p(x)\rho_x, \tag{4.102}$$

where the condition $\tau \geq p(x)\rho_x$ is the same as $\tau - p(x)\rho_x \geq 0$ (the operator $\tau - p(x)\rho_x$ is a positive operator). Show that $\text{Tr}\{\tau\}$ is an upper bound on the expected success probability of the POVM. After doing so, consider the case of encoding n bits into a d-dimensional subspace. By choosing states uniformly at random (in the case of the ensemble $\{2^{-n}, \rho_i\}_{i\in\{0,1\}^n}$), show that the expected success probability is bounded above by $d\ 2^{-n}$. Thus, it is not possible to store more than n classical bits in n qubits and have a perfect success probability of retrieval (this is a simplified version of the Holevo bound, about which we will learn more in Chapter 11).

4.3 Composite Noisy Quantum Systems

We are again interested in the behavior of two or more quantum systems when we join them together. Some of the most exotic, truly "quantum" behavior occurs in joint quantum systems, and we observe a marked departure from the classical world.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

4.3.1 Independent Ensembles

Let us first suppose that we have two independent ensembles for quantum systems A and B. The first quantum system belongs to Alice and the second quantum system belongs to Bob, and they may or may not be spatially separated. Let $\{p_X(x), |\psi_x\rangle\}$ be the ensemble for the system A and let $\{p_Y(y), |\phi_y\rangle\}$ be the ensemble for the system B. Suppose for now that the state on system A is $|\psi_x\rangle$ for some x and the state on system B is $|\phi_y\rangle$ for some y. Then, using the composite system postulate of the noiseless quantum theory, the joint state for a given x and y is $|\psi_x\rangle \otimes |\phi_y\rangle$. The density operator for the joint quantum system is the expectation of the states $|\psi_x\rangle \otimes |\phi_y\rangle$ with respect to the random variables X and Y that describe the individual ensembles:

$$\mathbb{E}_{X,Y}\{(|\psi_X\rangle \otimes |\phi_Y\rangle)(\langle\psi_X| \otimes \langle\phi_Y|)\}. \tag{4.103}$$

The above expression is equal to the following one:

$$\mathbb{E}_{X,Y}\{|\psi_X\rangle\langle\psi_X| \otimes |\phi_Y\rangle\langle\phi_Y|\}, \tag{4.104}$$

because $(|\psi_x\rangle \otimes |\phi_y\rangle)(\langle\psi_x| \otimes \langle\phi_y|) = |\psi_x\rangle\langle\psi_x| \otimes |\phi_y\rangle\langle\phi_y|$. We then explicitly write out the expectation as a sum over probabilities:

$$\sum_{x,y} p_X(x)p_Y(y)|\psi_x\rangle\langle\psi_x| \otimes |\phi_y\rangle\langle\phi_y|. \tag{4.105}$$

We can distribute the probabilities and the sum because the tensor product obeys a distributive property:

$$\sum_x p_X(x)|\psi_x\rangle\langle\psi_x| \otimes \sum_y p_Y(y)|\phi_y\rangle\langle\phi_y|. \tag{4.106}$$

The density operator for this ensemble admits the following simple form:

$$\rho \otimes \sigma, \tag{4.107}$$

where ρ is the density operator of the X ensemble and σ is the density operator of the Y ensemble. We can say that Alice's local density operator is ρ and Bob's local density operator is σ. We call a density operator of the above form a *product state*. We should expect the density operator to factorize as it does above because we assumed that the ensembles are independent. There is nothing much that distinguishes this situation from the classical world, except for the fact that the states in each respective ensemble may be non-orthogonal to other states in the same ensemble. But even here, there is some equivalent description of each ensemble in terms of an orthonormal basis so that there is really no difference between this description and a joint probability distribution that factors as two independent distributions.

Exercise 4.3.1 Show that the purity $P(\rho^A)$ is equal to the following expression

$$P(\rho^A) = \text{Tr}\left\{\left(\rho^A \otimes \rho^{A'}\right)F^{AA'}\right\} \tag{4.108}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where system A' has a Hilbert space structure isomorphic to that of system A and $F^{AA'}$ is the swap operator that has the following action on kets in A and A':

$$\forall x, y \qquad F^{AA'}|x\rangle^A|y\rangle^{A'} = |y\rangle^A|x\rangle^{A'}. \tag{4.109}$$

(Hint: First show that $\text{Tr}\{f(\rho^A)\} = \text{Tr}\{(f(\rho^A) \otimes I^{A'})F^{AA'}\}$ for any function f on the operators in system A.)

4.3.2 Separable States

Let us now consider two systems A and B whose corresponding ensembles are correlated. We describe this correlated ensemble as the joint ensemble

$$\{p_X(x), |\psi_x\rangle \otimes |\phi_x\rangle\}. \tag{4.110}$$

It is straightforward to verify that the density operator of this correlated ensemble has the following form:

$$\mathbb{E}_X\{(|\psi_X\rangle \otimes |\phi_X\rangle)(\langle\psi_X| \otimes \langle\phi_X|)\} = \sum_x p_X(x)|\psi_x\rangle\langle\psi_x| \otimes |\phi_x\rangle\langle\phi_x|. \tag{4.111}$$

The above state is a *separable* state. The term "separable" implies that there is no quantum entanglement in the above state, i.e., there is a completely classical procedure that prepares the above state. By ignoring Bob's system, Alice's local density operator is of the form

$$\mathbb{E}_X\{|\psi_X\rangle\langle\psi_X|\} = \sum_x p_X(x)|\psi_x\rangle\langle\psi_x|, \tag{4.112}$$

and similarly, Bob's local density operator is

$$\mathbb{E}_X\{|\phi_X\rangle\langle\phi_X|\} = \sum_x p_X(x)|\phi_x\rangle\langle\phi_x|. \tag{4.113}$$

We can generalize this classical preparation procedure one step further, using an idea similar to the "ensemble of ensembles" idea in Section 4.1.2. Let us suppose that we first generate a random variable Z according to some distribution $p_Z(z)$. We then generate two other ensembles, conditional on the value of the random variable Z. Let $\{p_{X|Z}(x|z), |\psi_{x,z}\rangle\}$ be the first ensemble and let $\{p_{Y|Z}(y|z), |\phi_{y,z}\rangle\}$ be the second ensemble, where the random variables X and Y are independent when conditioned on Z. Let us label the density operators of the first and second ensembles when conditioned on a particular realization z by ρ_z and σ_z, respectively. It is then straightforward to verify that the density operator of an ensemble created from this classical preparation procedure has the following form:

$$\mathbb{E}_{X,Y,Z}\{(|\psi_{X,Z}\rangle \otimes |\phi_{Y,Z}\rangle)(\langle\psi_{X,Z}| \otimes \langle\phi_{Y,Z}|)\} = \sum_z p_Z(z)\rho_z \otimes \sigma_z. \tag{4.114}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 4.3.2 By ignoring Bob's system, we can determine Alice's local density operator. Show that

$$\mathbb{E}_{X,Y,Z}\{|\psi_{X,Z}\rangle\langle\psi_{X,Z}|\} = \sum_z p_Z(z)\rho_z, \tag{4.115}$$

so that the above expression is the density operator for Alice. It similarly follows that the local density operator for Bob is

$$\mathbb{E}_{X,Y,Z}\{|\phi_{Y,Z}\rangle\langle\phi_{Y,Z}|\} = \sum_z p_Z(z)\sigma_z. \tag{4.116}$$

The density operator in (4.114) is the most general form of a separable state because the above procedure is the most general classical preparation procedure that produces classical correlations (we could generalize further with more ensembles of ensembles, but they would ultimately lead to this form because the set of separable states is a convex set). A state is *entangled* if we cannot write it in the form in (4.114), as a convex combination of product states.

Exercise 4.3.3 Show that we can always write a separable state as a convex combination of pure product states:

$$\sum_z p_Z(z)|\phi_z\rangle\langle\phi_z| \otimes |\psi_z\rangle\langle\psi_z|, \tag{4.117}$$

by manipulating the general form in (4.114).

4.3.3 Local Density Operator

A First Example

Consider the entangled Bell state $|\Phi^+\rangle^{AB}$ shared on systems A and B. In the above analyses, we were concerned with determining a local density operator description for both Alice and Bob. Now, we are curious if it is possible to determine such a local density operator description for Alice and Bob with respect to the state $|\Phi^+\rangle^{AB}$.

As a first approach to this issue, recall that the density operator description arises from its usefulness in determining the probabilities of the outcomes of a particular measurement. We say that the density operator is "the state" of the system merely because it is a mathematical representation that allows us to compute the probabilities resulting from a physical measurement. So, if we would like to determine a "local density operator," such a local density operator should predict the result of a local measurement.

Let us consider a local measurement with measurement operators $\{M_m\}_m$ that Alice can perform on her system. The global measurement operators for this local measurement are $\{M_m^A \otimes I^B\}_m$ because nothing (the identity) happens to Bob's system. The probability of

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

obtaining result m is

$$\left\langle \Phi^+ \right| M_m^A \otimes I^B \left| \Phi^+ \right\rangle = \frac{1}{2} \sum_{i,j=0}^{1} \langle ii | M_m^A \otimes I^B | jj \rangle \tag{4.118}$$

$$= \frac{1}{2} \sum_{i,j=0}^{1} \langle i | M_m^A | j \rangle \langle i | j \rangle \tag{4.119}$$

$$= \frac{1}{2} \big(\langle 0 | M_m^A | 0 \rangle + \langle 1 | M_m^A | 1 \rangle \big) \tag{4.120}$$

$$= \frac{1}{2} \Big(\text{Tr} \Big\{ M_m^A |0\rangle\langle 0|^A \Big\} + \text{Tr} \Big\{ M_m^A |1\rangle\langle 1|^A \Big\} \Big) \tag{4.121}$$

$$= \text{Tr} \Big\{ M_m^A \frac{1}{2} \Big(|0\rangle\langle 0|^A + |1\rangle\langle 1|^A \Big) \Big\} \tag{4.122}$$

$$= \text{Tr} \big\{ M_m^A \pi^A \big\}. \tag{4.123}$$

The above steps follow by applying the rules of taking the inner product with respect to tensor product operators. The last line follows by recalling the definition of the maximally mixed state π in (4.35), where π here is a qubit maximally mixed state.

The above calculation demonstrates that we can predict the result of any local "Alice" measurement using the density operator π. Therefore, it is reasonable to say that Alice's local density operator is π, and we even go as far to say that her *local state* is π. A symmetric calculation shows that Bob's local state is also π.

This result concerning their local density operators may seem strange at first. The following global state gives equivalent predictions for local measurements:

$$\pi^A \otimes \pi^B. \tag{4.124}$$

Can we then conclude that an equivalent representation of the global state is the above state? Absolutely not. The global state $|\Phi^+\rangle^{AB}$ and the above state give drastically different predictions for global measurements. Exercise 4.3.5 below asks you to determine the probabilities for measuring the global operator $Z^A \otimes Z^B$ when the global state is $|\Phi^+\rangle^{AB}$ or $\pi^A \otimes \pi^B$, and the result is that the predictions are dramatically different.

Exercise 4.3.4 Show that the projection operators corresponding to a measurement of the observable $Z^A \otimes Z^B$ are as follows:

$$\Pi_{\text{even}} \equiv \frac{1}{2} \big(I^A \otimes I^B + Z^A \otimes Z^B \big) = |00\rangle\langle 00|^{AB} + |11\rangle\langle 11|^{AB}, \tag{4.125}$$

$$\Pi_{\text{odd}} \equiv \frac{1}{2} \big(I^A \otimes I^B - Z^A \otimes Z^B \big) = |01\rangle\langle 01|^{AB} + |10\rangle\langle 10|^{AB}. \tag{4.126}$$

This measurement is a parity measurement, where measurement operator Π_{even} coherently measures even parity and measurement operator Π_{odd} measures odd parity.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 4.3.5 Show that a parity measurement (defined in the previous exercise) of the state $|\Phi^+\rangle^{AB}$ returns an even parity result with probability one, and a parity measurement of the state $\pi^A \otimes \pi^B$ returns even or odd parity with equal probability. Thus, despite the fact that these states have the same local description, their global behavior is very different. Show that the same is true for the phase parity measurement, given by

$$\Pi_{X,\text{even}} \equiv \frac{1}{2}\big(I^A \otimes I^B + X^A \otimes X^B\big), \tag{4.127}$$

$$\Pi_{X,\text{odd}} \equiv \frac{1}{2}\big(I^A \otimes I^B - X^A \otimes X^B\big). \tag{4.128}$$

Exercise 4.3.6 Show that the maximally correlated state $\overline{\Phi}^{AB}$, where

$$\overline{\Phi}^{AB} = \frac{1}{2}\Big(|00\rangle\langle 00|^{AB} + |11\rangle\langle 11|^{AB}\Big), \tag{4.129}$$

gives results for local measurements that are the same as those for the maximally entangled state $|\Phi^+\rangle^{AB}$. Show that the above parity measurements can distinguish these states.

Partial Trace

In general, we would like to determine a local density operator that predicts the outcomes of all local measurements without having to resort repeatedly to an analysis like that in (4.118-4.123). The general method for determining a local density operator is to employ the *partial trace operation*. For a simple state of the form

$$|x\rangle\langle x| \otimes |y\rangle\langle y|, \tag{4.130}$$

the partial trace is the following operation:

$$|x\rangle\langle x| \operatorname{Tr}\{|y\rangle\langle y|\} = |x\rangle\langle x|, \tag{4.131}$$

where we "trace out" the second system to determine the local density operator for the first. We define it mathematically as acting on any tensor product of rank-one operators (not necessarily corresponding to a state)

$$|x_1\rangle\langle x_2| \otimes |y_1\rangle\langle y_2|, \tag{4.132}$$

as follows:

$$\operatorname{Tr}_2\{|x_1\rangle\langle x_2| \otimes |y_1\rangle\langle y_2|\} \equiv |x_1\rangle\langle x_2| \operatorname{Tr}\{|y_1\rangle\langle y_2|\} \tag{4.133}$$

$$= |x_1\rangle\langle x_2| \langle y_1|y_2\rangle. \tag{4.134}$$

The subscript "2" of the trace operation indicates that the partial trace acts on the second system. It is a linear operation, much like the full trace is a linear operation.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 4.3.7 Show that the partial trace operation is equivalent to

$$\text{Tr}_B\Big\{|x_1\rangle\langle x_2|^A \otimes |y_1\rangle\langle y_2|^B\Big\} = \sum_i \langle i|^B \Big(|x_1\rangle\langle x_2|^A \otimes |y_1\rangle\langle y_2|^B\Big)|i\rangle^B, \tag{4.135}$$

for some orthonormal basis $\{|i\rangle^B\}$ on Bob's system.

The most general density operator on two systems A and B is some operator ρ^{AB} that is positive with unit trace. We can obtain the local density operator ρ^A from ρ^{AB} by tracing out the B system:

$$\rho^A = \text{Tr}_B\{\rho^{AB}\}. \tag{4.136}$$

In more detail, let us expand an arbitrary density operator ρ^{AB} with an orthonormal basis $\{|i\rangle^A \otimes |j\rangle^B\}_{i,j}$ for the bipartite (two-party) state:

$$\rho^{AB} = \sum_{i,j,k,l} \lambda_{i,j,k,l}(|i\rangle^A \otimes |j\rangle^B)(\langle k|^A \otimes \langle l|^B). \tag{4.137}$$

The coefficients $\lambda_{i,j,k,l}$ are the matrix elements of ρ^{AB} with respect to the basis $\{|i\rangle^A \otimes |j\rangle^B\}_{i,j}$, and they are subject to the constraint of positivity and unit trace for ρ^{AB} . We can rewrite the above operator as

$$\rho^{AB} = \sum_{i,j,k,l} \lambda_{i,j,k,l}|i\rangle\langle k|^A \otimes |j\rangle\langle l|^B. \tag{4.138}$$

We can now evaluate the partial trace:

$$\begin{aligned}
\rho^A &= \text{Tr}_B\left\{\sum_{i,j,k,l} \lambda_{i,j,k,l}|i\rangle\langle k|^A \otimes |j\rangle\langle l|^B\right\} && (4.139)\\
&= \sum_{i,j,k,l} \lambda_{i,j,k,l}\text{Tr}_B\Big\{|i\rangle\langle k|^A \otimes |j\rangle\langle l|^B\Big\} && (4.140)\\
&= \sum_{i,j,k,l} \lambda_{i,j,k,l}|i\rangle\langle k|^A\text{Tr}\Big\{|j\rangle\langle l|^B\Big\} && (4.141)\\
&= \sum_{i,j,k,l} \lambda_{i,j,k,l}|i\rangle\langle k|^A\langle j|l\rangle && (4.142)\\
&= \sum_{i,j,k} \lambda_{i,j,k,j}|i\rangle\langle k|^A && (4.143)\\
&= \sum_{i,k}\left(\sum_j \lambda_{i,j,k,j}\right)|i\rangle\langle k|^A. && (4.144)
\end{aligned}$$

The second equality exploits the linearity of the partial trace operation. The last equality explicitly shows how the partial trace operation earns its name—it is equivalent to performing a trace operation over the coefficients corresponding to Bob's system.

The next exercise asks you to verify that the operator ρ^A, as defined by the partial trace, predicts the results of a local measurement accurately and confirms the role of ρ^A as a local density operator.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 4.3.8 Suppose Alice and Bob share a quantum system in a state described by the density operator ρ^{AB}. Consider a local measurement, with measurement operators $\{M_m\}_m$, that Alice may perform on her system. The global measurement operators are thus $\{M_m^A \otimes I^B\}_m$. Show that the probabilities predicted by the global density operator are the same as those predicted by the local density operator ρ^A where $\rho^A = \text{Tr}_B\{\rho^{AB}\}$:

$$\text{Tr}\{(M_m^A \otimes I^B)\rho^{AB}\} = \text{Tr}\{M_m^A \rho^A\}. \tag{4.145}$$

Thus, the predictions of the global noisy quantum theory are consistent with the predictions of the local noisy quantum theory.

Exercise 4.3.9 Verify that the partial trace of a product state gives one of the density operators in the product state:

$$\text{Tr}_2\{\rho \otimes \sigma\} = \rho. \tag{4.146}$$

This result is consistent with the observation near (4.107).

Exercise 4.3.10 Verify that the partial trace of a separable state gives the result in (4.115):

$$\text{Tr}_2\left\{\sum_z p_Z(z)\rho_z \otimes \sigma_z\right\} = \sum_z p_Z(z)\rho_z. \tag{4.147}$$

Exercise 4.3.11 Consider the following density operator that is formally analogous to a joint probability distribution $p_{X,Y}(x,y)$:

$$\rho = \sum_{x,y} p_{X,Y}(x,y)|x\rangle\langle x| \otimes |y\rangle\langle y|, \tag{4.148}$$

where the set of states $\{|x\rangle\}_x$ and $\{|y\rangle\}_y$ each form an orthonormal basis. Show that tracing out the second system is formally analogous to taking the marginal distribution $p_X(x) = \sum_y p_{X,Y}(x,y)$ of the joint distribution $p_{X,Y}(x,y)$. That is, we are left with a density operator of the form

$$\sum_x p_X(x)|x\rangle\langle x|. \tag{4.149}$$

Keep in mind that the partial trace is a generalization of the marginalization because it handles more exotic quantum states besides the above "classical" state.

Exercise 4.3.12 Show that the two partial traces in any order on a bipartite system are equivalent to a full trace:

$$\text{Tr}\{\rho^{AB}\} = \text{Tr}_A\{\text{Tr}_B\{\rho^{AB}\}\} = \text{Tr}_B\{\text{Tr}_A\{\rho^{AB}\}\}. \tag{4.150}$$

Exercise 4.3.13 Verify that Alice's local density operator does not change if Bob performs a unitary operator or a measurement where he does not inform her of the measurement result.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

4.3.4 Classical-Quantum Ensemble

We end our overview of composite noisy quantum systems by discussing one last type of joint ensemble: the *classical-quantum ensemble*. This ensemble is a generalization of the "ensemble of ensembles" from before.

Let us consider the following ensemble of density operators:

$$\{p_X(x), \rho_x\}_{x\in\mathcal{X}}. \tag{4.151}$$

The intuition here is that Alice prepares a quantum system in the state ρ_x with probability $p_X(x)$. She then passes this ensemble to Bob, and it is Bob's task to learn about it. He can learn about the ensemble if Alice prepares a large number of them in the same way.

There is generally a loss of the information in the random variable X once Alice has prepared this ensemble. Bob can only learn about the distribution of the random variable X if each density operator ρ_x is a pure state $|x\rangle\langle x|$ where the states $\{|x\rangle\}_{x\in\mathcal{X}}$ form an orthonormal basis. The resulting density operator would be

$$\rho = \sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x|. \tag{4.152}$$

Bob could then perform a measurement with measurement operators $\{|x\rangle\langle x|\}_{x\in\mathcal{X}}$, and learn about the distribution $p_X(x)$ with a large number of measurements.

In the general case, the density operators $\{\rho_x\}_{x\in\mathcal{X}}$ do not correspond to pure states, much less orthonormal ones, and there is no procedure for Bob to learn about random variable X. The density operator of the ensemble is

$$\rho = \sum_{x\in\mathcal{X}} p_X(x)\rho_x, \tag{4.153}$$

and the information about the distribution of random variable X becomes "mixed in" with the "mixedness" of the density operators ρ_x. There is then no measurement that Bob can perform on ρ that allows him to learn about the probability distribution of random variable X.

One solution to this issue is for Alice to prepare the following classical-quantum ensemble:

$$\left\{p_X(x), |x\rangle\langle x|^X \otimes \rho_x^A\right\}_{x\in\mathcal{X}}, \tag{4.154}$$

where we label the first system as X and the second as A. She simply correlates a state $|x\rangle$ with each density operator ρ_x, where the states $\{|x\rangle\}_{x\in\mathcal{X}}$ form an orthonormal basis. We call this ensemble the "classical-quantum" ensemble because the first system is classical and the second system is quantum. The density operator of the classical-quantum ensemble is a *classical-quantum state* ρ^{XA} where

$$\rho^{XA} \equiv \sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A. \tag{4.155}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

This "enlarged" ensemble lets Bob learn about random variable X while at the same time he can learn about the ensemble that Alice prepares. Bob can learn about the distribution of random variable X by performing a local measurement of the system X. He also can learn about the states ρ_x by performing a measurement on A and combining the result of this measurement with the result of the first measurement. The next exercises ask you to verify these statements.

Exercise 4.3.14 Show that a local measurement of system X reproduces the probability distribution $p_X(x)$. Use local measurement operators $\{|x\rangle\langle x|\}_{x\in\mathcal{X}}$ to show that

$$p_X(x) = \text{Tr}\Big\{\rho^{XA}\Big(|x\rangle\langle x|^X \otimes I^A\Big)\Big\}. \tag{4.156}$$

Exercise 4.3.15 Show that performing a measurement with measurement operators $\{M_m\}$ on system A is the same as performing a measurement of the ensemble in (4.151). Show that

$$\text{Tr}\{\rho M_m\} = \text{Tr}\big\{\rho^{XA}\big(I^X \otimes M_m^A\big)\big\}, \tag{4.157}$$

where ρ is defined in (4.153).

4.4 Noisy Evolution

The evolution of a quantum state is never perfect. In this section, we introduce noise as resulting from the loss of information about a quantum system. This loss of information is similar to the lack of information about the preparation of a quantum state, as we have seen in the previous section.

4.4.1 Noisy Evolution from a Random Unitary

We begin with an example, the quantum bit-flip channel. Suppose that we prepare a quantum state $|\psi\rangle$. For simplicity, let us suppose that we are able to prepare this state perfectly. Suppose that we send this qubit over a quantum bit-flip channel, i.e., the channel applies the X Pauli operator (bit-flip operator) with some probability p and applies the identity operator with probability $1-p$. We can describe the resulting state as the following ensemble:

$$\{\{p, X|\psi\rangle\}, \{1-p, |\psi\rangle\}\}. \tag{4.158}$$

The density operator of this ensemble is as follows:

$$pX|\psi\rangle\langle\psi|X^\dagger + (1-p)|\psi\rangle\langle\psi|. \tag{4.159}$$

We now generalize the above example by beginning with an ensemble

$$\{p_X(x), |\psi_x\rangle\}_{x\in\mathcal{X}} \tag{4.160}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

with density operator $\rho \equiv \sum_{x\in\mathcal{X}} p_X(x)|\psi_x\rangle\langle\psi_x|$ and apply the bit-flip channel to this ensemble. Given that the input state is $|\psi_x\rangle$, the resulting ensemble is as in (4.158) with $|\psi\rangle$ replaced by $|\psi_x\rangle$. The overall ensemble is then as follows:

$$\{\{p_X(x)p, X|\psi_x\rangle\}, \{p_X(x)(1-p), |\psi_x\rangle\}\}_{x\in\mathcal{X}}. \tag{4.161}$$

We can calculate the density operator of the above ensemble:

$$\sum_{x\in\mathcal{X}} p_X(x)pX|\psi_x\rangle\langle\psi_x|X^\dagger + p_X(x)(1-p)|\psi_x\rangle\langle\psi_x|, \tag{4.162}$$

and simplify the above density operator by employing the definition of ρ:

$$pX\rho X^\dagger + (1-p)\rho. \tag{4.163}$$

The above density operator is more "mixed" than the original density operator and we will make this statement more precise in Chapter 10, when we study entropy.

Random Unitaries

The generalization of the above discussion is to consider some ensemble of unitaries (a random unitary) $\{p(k), U_k\}$ that we can apply to an ensemble of states $\{p_X(x), |\psi_x\rangle\}_{x\in\mathcal{X}}$. It is straightforward to show that the resulting density operator is

$$\sum_k p(k)U_k\rho U_k^\dagger, \tag{4.164}$$

where ρ is the density operator of the ensemble of states.

4.4.2 Noisy Evolution as the Loss of a Measurement Outcome

We can also think about noise as arising from the loss of a measurement outcome. Suppose that we have an ensemble of states $\{p_X(x), |\psi_x\rangle\}_{x\in\mathcal{X}}$ and we perform a measurement with a set $\{M_k\}$ of measurement operators where $\sum_k M_k^\dagger M_k = I$. First let us suppose that we know that the state is $|\psi_x\rangle$. Then the probability of obtaining the measurement outcome k is $p_{K|X}(k|x)$ where

$$p_{K|X}(k|x) = \langle\psi_x|M_k^\dagger M_k|\psi_x\rangle, \tag{4.165}$$

and the post-measurement state is

$$\frac{M_k|\psi_x\rangle}{\sqrt{p_{K|X}(k|x)}}. \tag{4.166}$$

Let us now suppose that we lose track of the measurement outcome, or equivalently, someone else measures the system and does not inform us of the measurement outcome. The resulting ensemble description is then

$$\left\{p_{X|K}(x|k)p_K(k), M_k|\psi_x\rangle/\sqrt{p_{K|X}(k|x)}\right\}_{x\in\mathcal{X},k}. \tag{4.167}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

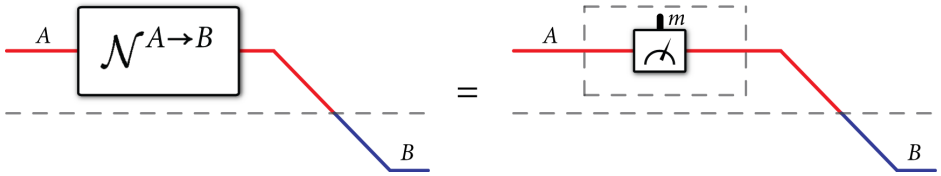

Figure 4.2: We use the diagram on the left to depict a noisy quantum channel $\mathcal{N}^{A\to B}$ that takes a quantum system A to a quantum system B. This quantum channel is equivalent to the diagram on the right, where some third party performs a measurement on the input system and does not inform the receiver of the measurement outcome.

The density operator of the ensemble is then

$$\begin{aligned}
&\sum_{x,k} p_{X|K}(x|k)p_K(k)\frac{M_k|\psi_x\rangle\langle\psi_x|M_k^\dagger}{p_{K|X}(k|x)} \\
&= \sum_{x,k} p_{K|X}(k|x)p_X(x)\frac{M_k|\psi_x\rangle\langle\psi_x|M_k^\dagger}{p_{K|X}(k|x)} && (4.168)\\
&= \sum_{x,k} p_X(x)M_k|\psi_x\rangle\langle\psi_x|M_k^\dagger && (4.169)\\
&= \sum_k M_k\rho M_k^\dagger. && (4.170)
\end{aligned}$$

We can thus write this evolution as a noisy map $\mathcal{N}(\rho)$ where

$$\mathcal{N}(\rho) \equiv \sum_k M_k\rho M_k^\dagger. \tag{4.171}$$

We derived the map in (4.171) from the perspective of the loss of a measurement outcome, but it in fact represents a general evolution of a density operator, and the operators M_k are known as the *Kraus operators*. We can represent all noisy evolutions in the form (4.171). The evolution of the density operator ρ is *trace-preserving* because the trace of the resulting density operator has unit trace:

$$\begin{aligned}
\text{Tr}\{\mathcal{N}(\rho)\} &= \text{Tr}\left\{\sum_k M_k\rho M_k^\dagger\right\} && (4.172)\\
&= \sum_k \text{Tr}\left\{M_k\rho M_k^\dagger\right\} && (4.173)\\
&= \sum_k \text{Tr}\left\{M_k^\dagger M_k\rho\right\} && (4.174)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$$= \text{Tr}\left\{ \sum_k M_k^\dagger M_k \rho \right\} \tag{4.175}$$

$$= \text{Tr}\{\rho\} \tag{4.176}$$

$$= 1. \tag{4.177}$$

There is another important condition that the map $\mathcal{N}(\rho)$ should satisfy: *complete positivity*. *Positivity* is a special case of complete positivity, and this condition is that the output $\mathcal{N}(\rho)$ is a positive operator whenever the input ρ is a positive operator. Positivity ensures that the noisy evolution produces a quantum state as an output whenever the input is a quantum state. Complete positivity is that the output of the tensor product map $\left(I^k \otimes \mathcal{N}\right)(\sigma)$ for any finite k is a positive operator whenever the input σ is a positive operator (this input operator now lives on a tensor-product Hilbert space). If the input dimension of the noisy map is d, then it is sufficient to consider $k = d$. Complete positivity makes good physical sense because we expect that the action of a noisy map on one system of a quantum state and the identity on the other part of that quantum state should produce as output another quantum state (which is a positive operator). The map $\mathcal{N}(\rho)$ is a completely positive trace-preserving map, and any physical evolution is such a map.

Exercise 4.4.1 Show that the evolution in (4.171) is positive, i.e., the evolution takes a positive density operator to a positive density operator.

Exercise 4.4.2 Show that the evolution in (4.171) is completely positive.

Exercise 4.4.3 Show that the evolution in (4.171) is linear:

$$\mathcal{N}\left(\sum_x p_X(x) \rho_x \right) = \sum_x p_X(x) \mathcal{N}(\rho_x), \tag{4.178}$$

for any probabilities $p_X(x)$ and density operators ρ_x.

Unitary evolution is a special case of the evolution in (4.171). We can think of it merely as some measurement where we always know the measurement outcome. That is, it is a measurement with one operator U in the set of measurement operators and it satisfies the completeness condition because $U^\dagger U = I$.

A completely positive trace-preserving map is the mathematical model that we use for a quantum channel in quantum Shannon theory because it represents the most general noisy evolution of a quantum state. This evolution is a generalization of the conditional probability distribution noise model of classical information theory. To see this, suppose that the input density operator ρ is of the following form:

$$\rho = \sum_x p_X(x) |x\rangle\langle x|, \tag{4.179}$$

where $\{|x\rangle\}$ is some orthonormal basis. We consider a channel $\mathcal{N}$ with the Kraus operators

$$\left\{ \sqrt{p_{Y|X}(y|x)} |y\rangle\langle x| \right\}_{x,y}, \tag{4.180}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $|y\rangle$ and $|x\rangle$ are part of the same basis. Evolution according to this map is then as follows:

$$\mathcal{N}(\rho) = \sum_{x,y} \sqrt{p_{Y|X}(y|x)}|y\rangle\langle x|\left(\sum_{x'} p_X(x')|x'\rangle\langle x'|\right)\sqrt{p_{Y|X}(y|x)}|x\rangle\langle y| \tag{4.181}$$

$$= \sum_{x,y,x'} p_{Y|X}(y|x)p_X(x')|\langle x'|x\rangle|^2 \ |y\rangle\langle y| \tag{4.182}$$

$$= \sum_{x,y} p_{Y|X}(y|x)p_X(x) \ |y\rangle\langle y| \tag{4.183}$$

$$= \sum_{y}\left(\sum_{x} p_{Y|X}(y|x)p_X(x)\right) \ |y\rangle\langle y|. \tag{4.184}$$

Thus, the evolution is the same that a noisy classical channel $p_{Y|X}(y|x)$ would enact on a probability distribution $p_X(x)$ by taking it to

$$p_Y(y) = \sum_x p_{Y|X}(y|x)p_X(x) \tag{4.185}$$

at the output.

4.4.3 Noisy Evolution from a Unitary Interaction

There is another perspective on quantum noise that is useful to consider. It is equivalent to the perspective given in Chapter 5 when we discuss isometric evolution. Suppose that a quantum system A begins in the state ρ^A and that there is an environment system E in a pure state $|0\rangle^E$. So the initial state of the joint system AE is

$$\rho^A \otimes |0\rangle\langle 0|^E. \tag{4.186}$$

Suppose that these two systems interact according to some unitary operator U^{AE} acting on the tensor-product space of A and E. If we are only interested in the state σ^A of the system A after the interaction, then we find it by taking the partial trace over the environment E:

$$\sigma^A = \text{Tr}_E\left\{U^{AE}\left(\rho^A \otimes |0\rangle\langle 0|^E\right)\left(U^{AE}\right)^\dagger\right\}. \tag{4.187}$$

This evolution is equivalent to that of a completely-positive, trace-preserving map with Kraus operators $\{B_i \equiv \langle i|^E U^{AE}|0\rangle^E\}_i$. This follows easily because we can take the partial trace

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

with respect to an orthonormal basis $\{|i\rangle^E\}$ for the environment:

$$\begin{aligned}
&\text{Tr}_E\Big\{U^{AE}\Big(\rho^A \otimes |0\rangle\langle 0|^E\Big)\big(U^{AE}\big)^\dagger\Big\} \\
&= \sum_i \langle i|^E U^{AE}\Big(\rho^A \otimes |0\rangle\langle 0|^E\Big)\big(U^{AE}\big)^\dagger |i\rangle^E && (4.188)\\
&= \sum_i \langle i|^E U^{AE}|0\rangle^E \rho^A \langle 0|^E \big(U^{AE}\big)^\dagger |i\rangle^E && (4.189)\\
&= \sum_i B_i \rho B_i^\dagger. && (4.190)
\end{aligned}$$

That the operators $\{B_i\}$ are a legitimate set of Kraus operators satisfying $\sum_i B_i^\dagger B_i = I^A$ follows from the unitarity of U^{AE} and the orthonormality of the basis $\{|i\rangle^E\}$:

$$\begin{aligned}
\sum_i B_i^\dagger B_i &= \sum_i \langle 0|^E \big(U^{AE}\big)^\dagger |i\rangle^E \langle i|^E U^{AE}|0\rangle^E && (4.191)\\
&= \langle 0|^E \big(U^{AE}\big)^\dagger \sum_i |i\rangle\langle i|^E U^{AE}|0\rangle^E && (4.192)\\
&= \langle 0|^E \big(U^{AE}\big)^\dagger U^{AE}|0\rangle^E && (4.193)\\
&= \langle 0|^E I^A \otimes I^E |0\rangle^E && (4.194)\\
&= I^A. && (4.195)
\end{aligned}$$

4.4.4 Unique Specification of a Noisy Channel

Consider a given noisy quantum channel $\mathcal{N}$ with Kraus representation

$$\mathcal{N}(\rho) = \sum_j A_j \rho A_j^\dagger. \tag{4.196}$$

We can also uniquely specify $\mathcal{N}$ by its action on an operator of the form $|i\rangle\langle j|$ where $\{|i\rangle\}$ is some orthonormal basis:

$$N_{ij} \equiv \mathcal{N}(|i\rangle\langle j|). \tag{4.197}$$

We can figure out how the channel $\mathcal{N}$ would act on any density operator if we know how it acts on $|i\rangle\langle j|$ for all i and j. Thus, two channels $\mathcal{N}$ and $\mathcal{M}$ are equivalent if they have the same effect on all operators of the form $|i\rangle\langle j|$:

$$\mathcal{N} = \mathcal{M} \quad \Leftrightarrow \quad \forall i,j \quad \mathcal{N}(|i\rangle\langle j|) = \mathcal{M}(|i\rangle\langle j|). \tag{4.198}$$

Let us now consider a maximally entangled qudit state $|\Phi\rangle^{AB}$ where

$$|\Phi\rangle^{AB} = \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |i\rangle^A |i\rangle^B, \tag{4.199}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and d is the dimension of each system A and B. The density operator Φ^{AB} of such a state is as follows:

$$\Phi^{AB} = \frac{1}{d}\sum_{i,j=0}^{d-1} |i\rangle\langle j|^A \otimes |i\rangle\langle j|^B. \tag{4.200}$$

Let us now send the A system of Φ^{AB} through the noisy quantum channel $\mathcal{N}$:

$$\left(\mathcal{N}^A \otimes I^B\right)\left(\Phi^{AB}\right) = \frac{1}{d}\sum_{i,j=0}^{d-1} \mathcal{N}^A(|i\rangle\langle j|^A) \otimes |i\rangle\langle j|^B. \tag{4.201}$$

The resulting state completely characterizes the noisy channel $\mathcal{N}$ because the following map translates between the state in (4.201) and the operators N_{ij} in (4.197):

$$d\langle i'|\left(\mathcal{N}^A \otimes I^B\right)\left(\Phi^{AB}\right)|j'\rangle^B = N_{ij}. \tag{4.202}$$

Thus, we can completely characterize a noisy map by determining the quantum state resulting from sending half of a maximally entangled state through it, and the following condition is sufficient for any two noisy channels to be equivalent:

$$\mathcal{N} = \mathcal{M} \quad \Leftrightarrow \quad \left(\mathcal{N}^A \otimes I^B\right)\left(\Phi^{AB}\right) = \left(\mathcal{M}^A \otimes I^B\right)\left(\Phi^{AB}\right). \tag{4.203}$$

It is equivalent to the condition in (4.198).

4.4.5 Concatenation of Noisy Maps

A quantum state may undergo not just one type of noisy evolution—it can of course undergo one noisy quantum channel followed by another noisy quantum channel. Let $\mathcal{N}_1$ denote a first noisy evolution and let $\mathcal{N}_2$ denote a second noisy evolution. Suppose that the Kraus operators of $\mathcal{N}_1$ are $\{A_k\}$ and the Kraus operators of $\mathcal{N}_2$ are $\{B_k\}$. It is straightforward to define the concatenation $\mathcal{N}_2 \circ \mathcal{N}_1$ of these two maps. Consider that the output of the first map is

$$\mathcal{N}_1(\rho) \equiv \sum_k A_k \rho A_k^\dagger, \tag{4.204}$$

for some input density operator ρ. The output of the concatenation map $\mathcal{N}_2 \circ \mathcal{N}_1$ is then

$$(\mathcal{N}_2 \circ \mathcal{N}_1)(\rho) = \sum_k B_k \mathcal{N}_1(\rho) B_k^\dagger = \sum_{k,k'} B_k A_{k'} \rho A_{k'}^\dagger B_k^\dagger. \tag{4.205}$$

It is clear that the Kraus operators of the concatenation map are $\{B_k A_{k'}\}_{k,k'}$.

4.4.6 Important Examples of Noisy Evolution

This section discusses some of the most important examples of noisy evolutions that we will consider in this book. Throughout this book, we will be considering the information-carrying ability of these various channels. They will provide some useful, "hands on" insight into quantum Shannon theory.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Dephasing Channel

We have already given the example of a noisy quantum bit flip channel in Section 4.4.1. Another important example is a bit flip in the conjugate basis, or equivalently, a *phase flip channel*. This channel acts as follows on any given density operator:

$$\rho \rightarrow (1-p)\rho + pZ\rho Z. \tag{4.206}$$

It is also known as the *dephasing channel*.

For $p = 1/2$, the action of the dephasing channel on a given quantum state is equivalent to the action of measuring the qubit in the computational basis and forgetting the result of the measurement. We make this idea more clear with an example. First, suppose that we have a qubit

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \tag{4.207}$$

and we measure it in the computational basis. Then the postulates of quantum theory state that the qubit becomes $|0\rangle$ with probability $|\alpha|^2$ and it becomes $|1\rangle$ with probability $|\beta|^2$. Suppose that we forget the measurement outcome, or alternatively, we do not have access to it. Then our best description of the qubit is with the following ensemble:

$$\left\{\left\{|\alpha|^2, |0\rangle\right\}, \left\{|\beta|^2, |1\rangle\right\}\right\}. \tag{4.208}$$

The density operator of this ensemble is

$$|\alpha|^2|0\rangle\langle 0| + |\beta|^2|1\rangle\langle 1|. \tag{4.209}$$

Now let us check if the dephasing channel gives the same behavior as the forgetful measurement above. We can consider the qubit as being an ensemble $\{1, |\psi\rangle\}$, i.e., the state is certain to be $|\psi\rangle$. The density operator of the ensemble is then ρ where

$$\rho = |\alpha|^2|0\rangle\langle 0| + \alpha\beta^*|0\rangle\langle 1| + \alpha^*\beta|1\rangle\langle 0| + |\beta|^2|1\rangle\langle 1|. \tag{4.210}$$

If we act on the density operator ρ with the dephasing channel with $p = 1/2$, then it preserves the density operator with probability $1/2$ and phase flips the qubit with probability $1/2$:

$$\begin{aligned}
&\frac{1}{2}\rho + \frac{1}{2}Z\rho Z \\
&= \frac{1}{2}\left(|\alpha|^2|0\rangle\langle 0| + \alpha\beta^*|0\rangle\langle 1| + \alpha^*\beta|1\rangle\langle 0| + |\beta|^2|1\rangle\langle 1|\right) + \\
&\quad \frac{1}{2}\left(|\alpha|^2|0\rangle\langle 0| - \alpha\beta^*|0\rangle\langle 1| - \alpha^*\beta|1\rangle\langle 0| + |\beta|^2|1\rangle\langle 1|\right) &(4.211)\\
&= |\alpha|^2|0\rangle\langle 0| + |\beta|^2|1\rangle\langle 1|. &(4.212)
\end{aligned}$$

The dephasing channel nullifies the off-diagonal terms in the density operator with respect to the computational basis. The resulting density operator description is the same as what we found for the forgetful measurement.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 4.4.4 Verify that the action of the dephasing channel on the Bloch vector is

$$\frac{1}{2}(I + r_x X + r_y Y + r_z Z) \rightarrow \frac{1}{2}(I + (1-2p) r_x X + (1-2p) r_y Y + r_z Z), \quad (4.213)$$

so that the channel preserves any component of the Bloch vector in the Z direction, while shrinking any component in the X or Y direction.

Pauli Channel

A Pauli channel is a generalization of the above dephasing channel and the bit flip channel. It simply applies a random Pauli operator according to a probability distribution. The map for a qubit Pauli channel is

$$\rho \rightarrow \sum_{i,j=0}^{1} p(i,j) Z^i X^j \rho X^j Z^i. \quad (4.214)$$

The generalization of this channel to qudits is straightforward. We simply replace the Pauli operators with the Heisenberg-Weyl operators. The Pauli qudit channel is

$$\rho \rightarrow \sum_{i,j=0}^{d-1} p(i,j) Z(i) X(j) \rho X^\dagger(j) Z^\dagger(i). \quad (4.215)$$

These channels are important in the study of quantum key distribution (QKD) because an eavesdropper induces such a channel in a QKD protocol.

Exercise 4.4.5 We can write a Pauli channel as

$$\rho \rightarrow p_I \rho + p_X X \rho X + p_Y Y \rho Y + p_Z Z \rho Z. \quad (4.216)$$

Verify that the action of the Pauli channel on the Bloch vector is

$$(r_x, r_y, r_z) \rightarrow ((p_I + p_X - p_Y - p_Z) r_x,\ (p_I + p_Y - p_X - p_Z) r_y,\ (p_I + p_Z - p_X - p_Y) r_z). \quad (4.217)$$

Depolarizing Channel

The depolarizing channel is a "worst-case scenario" channel. It assumes that we just completely lose the input qubit with some probability, i.e., it replaces the lost qubit with the maximally mixed state. The map for the depolarizing channel is

$$\rho \rightarrow (1-p)\rho + p\pi, \quad (4.218)$$

where π is the maximally mixed state: $\pi = I/2$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Most of the time, this channel is too pessimistic. Usually, we can learn something about the physical nature of the channel by some estimation process. We should only consider using the depolarizing channel as a model if we have little to no information about the actual physical channel.

Exercise 4.4.6 (Pauli Twirl) Show that randomly applying the Pauli operators I, X, Y, Z with uniform probability to any density operator gives the maximally mixed state:

$$\frac{1}{4}\rho + \frac{1}{4}X\rho X + \frac{1}{4}Y\rho Y + \frac{1}{4}Z\rho Z = \pi. \tag{4.219}$$

(Hint: Represent the density operator as $\rho = (I + r_x X + r_y Y + r_z Z)/2$ and apply the commutation rules of the Pauli operators.) This is known as the "twirling" operation.

Exercise 4.4.7 Show that we can rewrite the depolarizing channel as the following Pauli channel:

$$\rho \to (1 - 3p/4)\rho + p\left(\frac{1}{4}X\rho X + \frac{1}{4}Y\rho Y + \frac{1}{4}Z\rho Z\right). \tag{4.220}$$

Exercise 4.4.8 Show that the action of a depolarizing channel on the Bloch vector is

$$(r_x, r_y, r_z) \to ((1-p)r_x,\ (1-p)r_y,\ (1-p)r_z). \tag{4.221}$$

Thus, it uniformly shrinks the Bloch vector to become closer to the maximally mixed state.

The generalization of the depolarizing channel to qudits is again straightforward. It is the same as the map in (4.218), with the exception that the density operators ρ and π are qudit density operators.

Exercise 4.4.9 (Qudit Twirl) Show that randomly applying the Heisenberg-Weyl operators

$$\{X(i)Z(j)\}_{i,j\in\{0,\ldots,d-1\}} \tag{4.222}$$

with uniform probability to any qudit density operator gives the maximally mixed state π:

$$\frac{1}{d^2}\sum_{i,j=0}^{d-1} X(i)Z(j)\rho Z^\dagger(j)X^\dagger(i) = \pi. \tag{4.223}$$

(Hint: You can do the full calculation, or you can decompose this channel into the composition of two completely dephasing channels where the first is a dephasing in the computational basis and the next is a dephasing in the conjugate basis).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Amplitude Damping Channel

The amplitude damping channel is a first-order approximation to a noisy evolution that occurs in many physical systems ranging from optical systems to chains of spin-1/2 particles to spontaneous emission of a photon from an atom.

In order to motivate this channel, we give a physical interpretation to our computational basis states. Let us think of the $|0\rangle$ state as the ground state of a two-level atom and let us think of the state $|1\rangle$ as the excited state of the atom. Spontaneous emission is a process that tends to decay the atom from its excited state to its ground state, even if the atom is in a superposition of the ground and excited states. Let the parameter γ denote the probability of decay so that $0 \leq \gamma \leq 1$. One Kraus operator that captures the decaying behavior is

$$A_0 = \sqrt{\gamma}|0\rangle\langle 1|. \tag{4.224}$$

The operator A_0 annihilates the ground state:

$$A_0|0\rangle\langle 0|A_0^\dagger = 0, \tag{4.225}$$

and it decays the excited state to the ground state:

$$A_0|1\rangle\langle 1|A_0^\dagger = \gamma|0\rangle\langle 0|. \tag{4.226}$$

The Kraus operator A_0 alone does not specify a physical map because $A_0^\dagger A_0 = \gamma|1\rangle\langle 1|$ (recall that the Kraus operators of any channel should satisfy the condition $\sum_k A_k^\dagger A_k = I$). We can satisfy this condition by choosing another operator A_1 such that

$$\begin{aligned} A_1^\dagger A_1 &= I - A_0^\dagger A_0 && (4.227)\\ &= |0\rangle\langle 0| + (1-\gamma)|1\rangle\langle 1|. && (4.228) \end{aligned}$$

The following choice of A_1 satisfies the above condition:

$$A_1 \equiv |0\rangle\langle 0| + \sqrt{1-\gamma}|1\rangle\langle 1|. \tag{4.229}$$

Thus, the operators A_0 and A_1 are valid Kraus operators for the amplitude damping channel.

Exercise 4.4.10 Consider a single-qubit density operator with the following matrix representation with respect to the computational basis:

$$\rho = \begin{bmatrix} 1-p & \eta \\ \eta^* & p \end{bmatrix}, \tag{4.230}$$

where $0 \leq p \leq 1$ and η is some complex number. Show that applying the amplitude damping channel with parameter γ to a qubit with the above density operator gives a density operator with the following matrix representation:

$$\begin{bmatrix} 1-(1-\gamma)p & \sqrt{1-\gamma}\eta \\ \sqrt{1-\gamma}\eta^* & (1-\gamma)p \end{bmatrix}. \tag{4.231}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 4.4.11 Show that the amplitude damping channel obeys a composition rule. Consider an amplitude damping channel $\mathcal{N}_1$ with transmission parameter $(1-\gamma_1)$ and consider another amplitude damping channel $\mathcal{N}_2$ with transmission parameter $(1-\gamma_2)$. Show that the composition channel $\mathcal{N}_2 \circ \mathcal{N}_1$ is an amplitude damping channel with transmission parameter $(1-\gamma_1)(1-\gamma_2)$. (Note that the transmission parameter is equal to one minus the damping parameter.)

Erasure Channel

The erasure channel is another important channel in quantum Shannon theory. It admits a simple model and is amenable to relatively straightforward analysis when we later discuss its capacity. The erasure channel can serve as a simplified model of photon loss in optical systems.

We first recall the classical definition of an erasure channel. A classical erasure channel either transmits a bit with some probability $1-\varepsilon$ or replaces it with an erasure symbol e with some probability ε. The output alphabet contains one more symbol than the input alphabet, namely, the erasure symbol e.

The generalization of the classical erasure channel to the quantum world is straightforward. It implements the following map:

$$\rho \to (1-\varepsilon)\rho + \varepsilon |e\rangle\langle e|, \tag{4.232}$$

where $|e\rangle$ is some state that is not in the input Hilbert space, and thus is orthogonal to it. The output space of the erasure channel is larger than its input space by one dimension. The interpretation of the quantum erasure channel is similar to that for the classical erasure channel. It transmits a qubit with probability $1-\varepsilon$ and "erases" it (replaces it with an orthogonal erasure state) with probability ε.

Exercise 4.4.12 Show that the following operators are the Kraus operators for the quantum erasure channel:

$$\sqrt{1-\varepsilon}(|0\rangle^B\langle 0|^A + |1\rangle^B\langle 1|^A), \tag{4.233}$$

$$\sqrt{\varepsilon}|e\rangle^B\langle 0|^A, \tag{4.234}$$

$$\sqrt{\varepsilon}|e\rangle^B\langle 1|^A. \tag{4.235}$$

At the receiving end of the channel, a simple measurement can determine whether an erasure has occurred. We perform a measurement with measurement operators $\{\Pi_{\text{in}}, |e\rangle\langle e|\}$, where Π_{in} is the projector onto the input Hilbert space. This measurement has the benefit of detecting no more information than necessary. It merely detects whether an erasure occurs, and thus preserves the quantum information at the input if an erasure does not occur.

Classical-Quantum Channel

A classical-quantum channel is one that first measures the input state in a particular orthonormal basis and outputs a density operator conditional on the result of the measurement.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

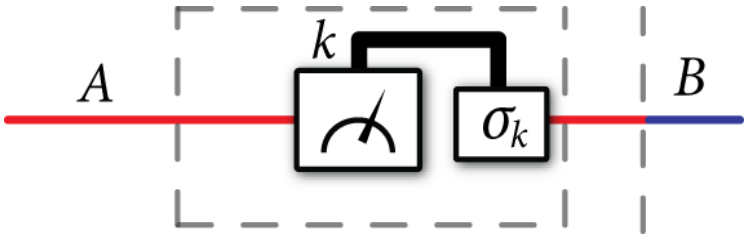

Figure 4.3: The above figure illustrates the internal workings of a classical-quantum channel. It first measures the input state in some basis $\{|k\rangle\}$ and outputs a quantum state σ_k conditional on the measurement outcome.

Suppose that the input to the channel is a density operator ρ. Suppose that $\{|k\rangle\}_k$ is an orthonormal basis for the Hilbert space on which the density operator ρ acts. The classical-quantum channel first measures the input state in the basis $\{|k\rangle\}$. Given that the result of the measurement is k, the post measurement state is

$$\frac{|k\rangle\langle k|\rho|k\rangle\langle k|}{\langle k|\rho|k\rangle}. \tag{4.236}$$

The classical-quantum channel correlates a density operator σ_k with the post-measurement state k:

$$\frac{|k\rangle\langle k|\rho|k\rangle\langle k|}{\langle k|\rho|k\rangle} \otimes \sigma_k. \tag{4.237}$$

This action leads to an ensemble:

$$\left\{\langle k|\rho|k\rangle, \frac{|k\rangle\langle k|\rho|k\rangle\langle k|}{\langle k|\rho|k\rangle} \otimes \sigma_k\right\}, \tag{4.238}$$

and the density operator of the ensemble is

$$\sum_k \langle k|\rho|k\rangle \frac{|k\rangle\langle k|\rho|k\rangle\langle k|}{\langle k|\rho|k\rangle} \otimes \sigma_k = \sum_k |k\rangle\langle k|\rho|k\rangle\langle k| \otimes \sigma_k. \tag{4.239}$$

The channel then only outputs the system on the right (tracing out the first system) so that the resulting channel is as follows:

$$\mathcal{N}(\rho) \equiv \sum_k \langle k|\rho|k\rangle \sigma_k. \tag{4.240}$$

Figure 4.3 depicts the behavior of the classical-quantum channel. This channel is also known as an entanglement-breaking channel, for reasons that become clear in the next exercise.

Exercise 4.4.13 Show that the classical-quantum channel is an *entanglement-breaking channel*—i.e., if we input the B system of an entangled state ψ^{AB}, then the resulting state on AB is no longer entangled.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can prove a more general structural theorem regarding entanglement-breaking channels by exploiting the observation in the above exercise.

Theorem 4.4.1. *An entanglement-breaking channel has a representation with Kraus operators that are unit rank.*

Proof. Consider that the output of an entanglement-breaking channel $\mathcal{N}_{\text{EB}}$ acting on half of a maximally entangled state is as follows:

$$\mathcal{N}_{\text{EB}}^{A\to B'}(\Phi^{BA}) = \sum_z p_Z(z)|\phi_z\rangle\langle\phi_z|^B \otimes |\psi_z\rangle\langle\psi_z|^{B'}. \tag{4.241}$$

This holds because the output of a channel is a separable state (it "breaks" entanglement), and it is always possible to find a representation of the separable state with pure states (see Exercise 4.3.3). Now consider constructing a channel $\mathcal{M}$ with the following unit-rank Kraus operators:

$$A_z \equiv \left\{\sqrt{d\ p_Z(z)}|\psi_z\rangle\langle\phi_z^*|\right\}_z, \tag{4.242}$$

where $|\phi_z^*\rangle$ is the state $|\phi_z\rangle$ with all of its elements conjugated. We should first verify that these Kraus operators form a valid channel, by checking that $\sum_z A_z^\dagger A_z = I$:

$$\sum_z A_z^\dagger A_z = \sum_z d\ p_Z(z)|\phi_z^*\rangle\langle\psi_z|\psi_z\rangle\langle\phi_z^*| \tag{4.243}$$

$$= d\ \sum_z p_Z(z)|\phi_z^*\rangle\langle\phi_z^*|. \tag{4.244}$$

Consider that

$$\text{Tr}_{B'}\left\{\mathcal{N}_{\text{EB}}^{A\to B'}(\Phi^{BA})\right\} = \pi^B \tag{4.245}$$

$$= \text{Tr}_{B'}\left\{\sum_z p_Z(z)|\phi_z\rangle\langle\phi_z|^B \otimes |\psi_z\rangle\langle\psi_z|^{B'}\right\} \tag{4.246}$$

$$= \sum_z p_Z(z)|\phi_z\rangle\langle\phi_z|^B, \tag{4.247}$$

where π^B is the maximally mixed state. Thus, it follows that $\mathcal{M}$ is a valid quantum channel because

$$d\ \sum_z p_Z(z)|\phi_z\rangle\langle\phi_z|^B = d\ \pi^B \tag{4.248}$$

$$= I^B \tag{4.249}$$

$$= \left(I^B\right)^* \tag{4.250}$$

$$= d\ \sum_z p_Z(z)|\phi_z^*\rangle\langle\phi_z^*| \tag{4.251}$$

$$= \sum_z A_z^\dagger A_z. \tag{4.252}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Now let us consider the action of the channel $\mathcal{M}$ on the maximally entangled state:

$$\mathcal{M}^{A\to B'}(\Phi^{BA}) \tag{4.253}$$

$$= \frac{1}{d}\sum_{z,i,j} |i\rangle\langle j|^B \otimes \sqrt{d\ p_Z(z)}|\psi_z\rangle\langle\phi_z^*||i\rangle\langle j||\phi_z^*\rangle\langle\psi_z|^{B'}\sqrt{d\ p_Z(z)} \tag{4.254}$$

$$= \sum_{z,i,j} p_Z(z)\ |i\rangle\langle j|^B \otimes \langle\phi_z^*|i\rangle\langle j|\phi_z^*\rangle\ |\psi_z\rangle\langle\psi_z|^{B'} \tag{4.255}$$

$$= \sum_{z,i,j} p_Z(z)\ |i\rangle\langle j|\phi_z^*\rangle\langle\phi_z^*|i\rangle\langle j|^B \otimes\ |\psi_z\rangle\langle\psi_z|^{B'} \tag{4.256}$$

$$= \sum_{z} p_Z(z)\ |\phi_z\rangle\langle\phi_z|^B \otimes\ |\psi_z\rangle\langle\psi_z|^{B'} \tag{4.257}$$

The last equality follows from recognizing $\sum_{i,j}|i\rangle\langle j|\cdot|i\rangle\langle j|$ as the transpose operation and noting that the transpose is equivalent to conjugation for an Hermitian operator $|\phi_z\rangle\langle\phi_z|$. Finally, since the action of both $\mathcal{N}_{\text{EB}}^{A\to B'}$ and $\mathcal{M}^{A\to B'}$ on the maximally entangled state is the same, we can conclude that the two channels are equivalent (see Section 4.4.4). Thus, $\mathcal{M}$ is a representation of the channel with unit-rank Kraus operators. □

4.4.7 Quantum Instrument

The description of a quantum channel with Kraus operators gives the most general evolution that a quantum state can undergo. We may want to specialize this definition somewhat for another scenario. Suppose that we would like to determine the most general evolution where the input is a quantum state and the output is both a quantum state and a classical variable. Such a scenario may arise in a case where Alice is trying to transmit both classical and quantum information, and Bob exploits a quantum instrument to decode both the classical and quantum systems. A *quantum instrument* gives such an evolution with a hybrid output.

Recall that we may view a noisy quantum channel as arising from the forgetting of a measurement outcome, as in (4.171). Let us now suppose that some third party performs a measurement with two outcomes j and k, but does not give us access to the measurement outcome j. Suppose that the measurement operators for this two-outcome measurement are $\{M_{j,k}\}_{j,k}$. Let us first suppose that the third party performs the measurement on a quantum system with density operator ρ and gives us both of the measurement outcomes. The post-measurement state in such a scenario is

$$\frac{M_{j,k}\rho M_{j,k}^\dagger}{p_{J,K}(j,k)}, \tag{4.258}$$

where the joint distribution of outcomes j and k are

$$p_{J,K}(j,k) = \text{Tr}\{M_{j,k}^\dagger M_{j,k}\rho\}. \tag{4.259}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can calculate the marginal distributions $p_J(j)$ and $p_K(k)$ according to the law of total probability:

$$p_J(j) = \sum_k p_{J,K}(j,k) = \sum_k \text{Tr}\{M_{j,k}^\dagger M_{j,k} \rho\}, \tag{4.260}$$

$$p_K(k) = \sum_j p_{J,K}(j,k) = \sum_j \text{Tr}\{M_{j,k}^\dagger M_{j,k} \rho\}. \tag{4.261}$$

Suppose the measuring device also places the classical outcomes in classical registers J and K, so that the post-measurement state is

$$\frac{M_{j,k} \rho M_{j,k}^\dagger}{p_{J,K}(j,k)} \otimes |j\rangle\langle j|^J \otimes |k\rangle\langle k|^K, \tag{4.262}$$

where the sets $\{|j\rangle\}$ and $\{|k\rangle\}$ form respective orthonormal bases. Such an operation is possible physically, and we could retrieve the classical information at some later point by performing a von Neumann measurement of the registers J and K. If we would like to determine the Kraus map for the overall quantum operation, we simply take the expectation over all measurement outcomes j and k:

$$\sum_{j,k} p_{J,K}(j,k) \left(\frac{M_{j,k} \rho M_{j,k}^\dagger}{p_{J,K}(j,k)} \right) \otimes |j\rangle\langle j|^J \otimes |k\rangle\langle k|^K$$
$$= \sum_{j,k} M_{j,k} \rho M_{j,k}^\dagger \otimes |j\rangle\langle j|^J \otimes |k\rangle\langle k|^K. \tag{4.263}$$

Let us now suppose that we do not have access to the measurement result k. This lack of access is equivalent to lacking access to classical register K. To determine the resulting state, we should trace out the classical register K. Our map then becomes

$$\sum_{j,k} M_{j,k} \rho M_{j,k}^\dagger \otimes |j\rangle\langle j|^J. \tag{4.264}$$

The above map corresponds to a quantum instrument, and is a general noisy quantum evolution that produces both a quantum output and a classical output. Figure 4.4 depicts a quantum instrument.

We can rewrite the above map more explicitly as follows:

$$\sum_j \left(\sum_k M_{j,k} \rho M_{j,k}^\dagger \right) \otimes |j\rangle\langle j|^J = \sum_j \mathcal{E}_j(\rho) \otimes |j\rangle\langle j|^J, \tag{4.265}$$

where we define

$$\mathcal{E}_j(\rho) \equiv \sum_k M_{j,k} \rho M_{j,k}^\dagger. \tag{4.266}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

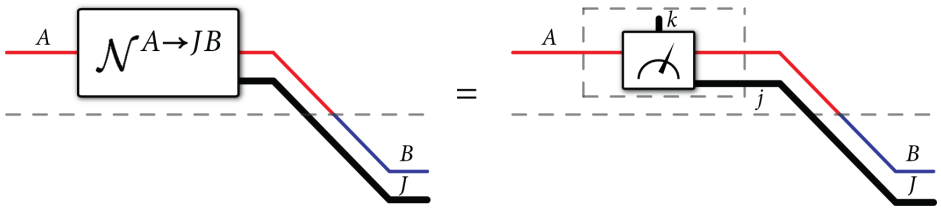

Figure 4.4: The figure on the left illustrates a quantum instrument, a general noisy evolution that produces both a quantum and classical output. The figure on the right illustrates the internal workings of a quantum instrument, showing that it results from having only partial access to a measurement outcome.

Each j-dependent map $\mathcal{E}_j(\rho)$ is a completely positive trace-*reducing* map because $\text{Tr}\{\mathcal{E}_j(\rho)\} \leq 1$. In fact, by examining the definition of $\mathcal{E}_j(\rho)$ and comparing to (4.260), it holds that

$$\text{Tr}\{\mathcal{E}_j(\rho)\} = p_J(j). \tag{4.267}$$

It is important to note that the probability $p_J(j)$ is dependent on the density operator ρ that is input to the instrument. We can determine the quantum output of the instrument by tracing over the classical register J. The resulting quantum output is then

$$\text{Tr}_J\left\{\sum_j \mathcal{E}_j(\rho) \otimes |j\rangle\langle j|^J\right\} = \sum_j \mathcal{E}_j(\rho). \tag{4.268}$$

The above "sum map" is a trace-preserving map because

$$\text{Tr}\left\{\sum_j \mathcal{E}_j(\rho)\right\} = \sum_j \text{Tr}\{\mathcal{E}_j(\rho)\} \tag{4.269}$$

$$= \sum_j p_J(j) \tag{4.270}$$

$$= 1, \tag{4.271}$$

where the last equality follows because the marginal probabilities $p_J(j)$ sum to one. The above points that we have mentioned are the most salient for the quantum instrument. We will exploit this type of evolution when we require a device that outputs both a classical and quantum system.

Exercise 4.4.14 Suppose that you have a set of completely-positive trace-*preserving* maps $\{\mathcal{E}_m\}$. Design a quantum instrument by modifying these maps in any way that you wish.

We should stress that a quantum instrument is more general than applying a mixture of CPTP maps to a quantum state. Suppose that we apply a mixture $\{\mathcal{N}_j\}$ of CPTP maps to a quantum state ρ, chosen according to a distribution $p_J(j)$. The resulting expected state is as follows:

$$\sum_j p_J(j)|j\rangle\langle j|^J \otimes \mathcal{N}_j(\rho). \tag{4.272}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

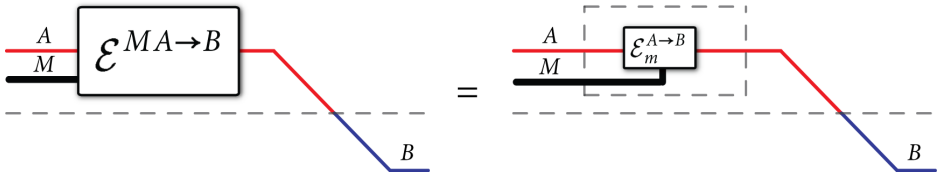

Figure 4.5: The figure on the left depicts a general operation, a conditional quantum encoder, that takes a classical system to a quantum system. The figure on the right depicts the inner workings of the conditional quantum encoder.

The probabilities $p_J(j)$ here are independent of the state ρ that is input to the mixture of CPTP maps, but this is not generally the case for a quantum instrument. There, the probabilities $p_J(j)$ can depend on the state ρ that is input—it may be beneficial then to write these probabilities as $p_J(j|\rho)$ because there is an implicit conditioning on the state that is input to the instrument.

4.4.8 Conditional Quantum Channel

We end this chapter by considering one final type of evolution. A *conditional quantum encoder* $\mathcal{E}^{MA\to B}$, or *conditional quantum channel* , is a collection $\{\mathcal{E}_m^{A\to B}\}_m$ of CPTP maps. Its inputs are a classical system M and a quantum system A and its output is a quantum system B. A conditional quantum encoder can function as an encoder of both classical and quantum information.

A classical-quantum state ρ^{MA}, where

$$\rho^{MA} \equiv \sum_m p(m)|m\rangle\langle m|^M \otimes \rho_m^A, \tag{4.273}$$

can act as an input to a conditional quantum encoder $\mathcal{E}^{MA\to B}$. The action of the conditional quantum encoder $\mathcal{E}^{MA\to B}$ on the classical-quantum state ρ^{MA} is as follows:

$$\mathcal{E}^{MA\to B}(\rho^{MA}) = \text{Tr}_M\left\{\sum_m p(m)|m\rangle\langle m|^M \otimes \mathcal{E}_m^{A\to B}(\rho_m^A)\right\}. \tag{4.274}$$

Figure 4.5 depicts the behavior of the conditional quantum encoder.

It is actually possible to write *any* quantum channel as a conditional quantum encoder when its input is a classical-quantum state. Indeed, consider any quantum channel $\mathcal{N}^{XA\to B}$ that has input systems X and A and output system B. Suppose the Kraus decomposition of this channel is as follows:

$$\mathcal{N}^{XA\to B}(\rho) \equiv \sum_j A_j \rho A_j^\dagger. \tag{4.275}$$

Suppose now that the input to the channel is the following classical-quantum state:

$$\sigma^{XA} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A. \tag{4.276}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Then the channel $\mathcal{N}^{XA\to B}$ acts as follows on the classical-quantum state σ^{XA}:

$$\mathcal{N}^{XA\to B}\left(\sigma^{XA}\right) = \sum_{j,x} A_j \left(p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A \right) A_j^\dagger. \tag{4.277}$$

Consider that a classical-quantum state admits the following matrix representation by exploiting the tensor product:

$$\sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A \tag{4.278}$$

$$= \begin{bmatrix} p_X(x_1)\rho_{x_1}^A & 0 & \cdots & 0 \\ 0 & p_X(x_2)\rho_{x_2}^A & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & p_X\left(x_{|\mathcal{X}|}\right)\rho_{x_{|\mathcal{X}|}}^A \end{bmatrix} \tag{4.279}$$

$$= \bigoplus_{x\in\mathcal{X}} p_X(x)\rho_x. \tag{4.280}$$

It is possible to specify a matrix representation for each Kraus operator A_j in terms of $|\mathcal{X}|$ block matrices:

$$A_j = \begin{bmatrix} A_{j,1} & A_{j,2} & \cdots & A_{j,|\mathcal{X}|} \end{bmatrix}. \tag{4.281}$$

Each operator $A_j \left(p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A \right) A_j^\dagger$ in the sum in (4.277) then takes the following form:

$$A_j \left(p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A \right) A_j^\dagger \tag{4.282}$$

$$= \begin{bmatrix} A_{j,1} & A_{j,2} & \cdots & A_{j,|\mathcal{X}|} \end{bmatrix} \begin{bmatrix} p_X(x_1)\rho_{x_1}^A & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & p_X\left(x_{|\mathcal{X}|}\right)\rho_{x_{|\mathcal{X}|}}^A \end{bmatrix} \begin{bmatrix} A_{j,1}^\dagger \\ A_{j,2}^\dagger \\ \vdots \\ A_{j,|\mathcal{X}|}^\dagger \end{bmatrix} \tag{4.283}$$

$$= \sum_{x\in|\mathcal{X}|} p_X(x) A_{j,x}\rho_x^A A_{j,x}^\dagger. \tag{4.284}$$

We can write the overall map as follows:

$$\mathcal{N}^{XA\to B}\left(\sigma^{XA}\right) = \sum_j \sum_{x\in\mathcal{X}} p_X(x) A_{j,x}\rho_x^A A_{j,x}^\dagger \tag{4.285}$$

$$= \sum_{x\in\mathcal{X}} p_X(x) \sum_j A_{j,x}\rho_x^A A_{j,x}^\dagger \tag{4.286}$$

$$= \sum_{x\in\mathcal{X}} p_X(x)\mathcal{N}_x^{A\to B}\left(\rho_x^A\right), \tag{4.287}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where we define each map $\mathcal{N}_x^{A\to B}$ as follows:

$$\mathcal{N}_x^{A\to B}(\rho_x^A) = \sum_j A_{j,x}\rho_x^A A_{j,x}^\dagger. \tag{4.288}$$

Thus, the action of any quantum channel on a classical-quantum state is the same as the action of the conditional quantum encoder.

Exercise 4.4.15 Show that the condition $\sum_j A_j^\dagger A_j = I$ implies the $|\mathcal{X}|$ conditions:

$$\forall x \in \mathcal{X} : \sum_j A_{j,x}^\dagger A_{j,x} = I. \tag{4.289}$$

4.5 Summary

We give a brief summary of the main results in this chapter. We derived all of these results from the noiseless quantum theory and an ensemble viewpoint. An alternate viewpoint is to say that the density operator is the state of the system and then give the postulates of quantum mechanics in terms of the density operator. Regardless of which viewpoint you view as more fundamental, they are consistent with each other in standard quantum mechanics.

The density operator ρ for an ensemble $\{p_X(x), |\psi_x\rangle\}$ is the following expectation:

$$\rho = \sum_x p_X(x)|\psi_x\rangle\langle\psi_x|. \tag{4.290}$$

The evolution of the density operator according to a unitary operator U is

$$\rho \to U\rho U^\dagger. \tag{4.291}$$

A measurement of the state according to a measurement $\{M_j\}$ where $\sum_j M_j^\dagger M_j = I$ leads to the following post-measurement state:

$$\rho \to \frac{M_j\rho M_j^\dagger}{p_J(j)}, \tag{4.292}$$

where the probability $p_J(j)$ for obtaining outcome j is

$$p_J(j) = \text{Tr}\Big\{M_j^\dagger M_j\rho\Big\}. \tag{4.293}$$

The most general noisy evolution that a quantum state can undergo is according to a completely-positive, trace-preserving map $\mathcal{N}(\rho)$ that we can write as follows:

$$\mathcal{N}(\rho) = \sum_j A_j\rho A_j^\dagger, \tag{4.294}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $\sum_j A_j^\dagger A_j = I$. A special case of this evolution is a quantum instrument. A quantum instrument has a quantum input and a classical and quantum output. The most general way to represent a quantum instrument is as follows:

$$\rho \to \sum_j \mathcal{E}_j(\rho) \otimes |j\rangle\langle j|^J, \tag{4.295}$$

where each map $\mathcal{E}_j$ is a completely-positive, trace-reducing map, where

$$\mathcal{E}_j = \sum_k A_{j,k} \rho A_{j,k}^\dagger, \tag{4.296}$$

and $\sum_{j,k} A_{j,k}^\dagger A_{j,k} = I$, so that the overall map is trace-preserving.

4.6 History and Further Reading

The book of Nielsen and Chuang gives an excellent introduction to noisy quantum channels [197]. Horodecki, Shor, and Ruskai introduced entanglement-breaking channels and proved several properties of them (e.g., the proof of Theorem 4.4.1) [150]. Davies and Lewis introduced the quantum instrument formalism [65], and Ozawa further elaborated it [201]. Grassl *et al.* introduced the quantum erasure channel and constructed some simple quantum error-correcting codes for it [114]. A discussion of the conditional quantum channel appears in Yard's thesis [266].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 5

The Purified Quantum Theory

The final chapter of our development of the quantum theory gives perhaps the most powerful viewpoint, by providing a mathematical tool, the purification theorem, which offers a completely different way of thinking about noise in quantum systems. This theorem shows that our lack of information about a set of quantum states can arise from entanglement with another system to which we do not have access. The system to which we do not have access is known as a *purification*. In this purified view of the quantum theory, noisy evolution arises from the interaction of a quantum system with its environment. The interaction of a quantum system with its environment leads to correlations between the quantum system and its environment, and this interaction leads to a loss of information because we cannot access the environment. The environment is thus the purification of the output of the noisy quantum channel.

In Chapter 3, we introduced the noiseless quantum theory. The noiseless quantum theory is a useful theory to learn so that we can begin to grasp an intuition for some uniquely quantum behavior, but it is an idealized model of quantum information processing. In Chapter 4, we introduced the noisy quantum theory as a generalization of the noiseless quantum theory. The noisy quantum theory can describe the behavior of imperfect quantum systems that are subject to noise.

In this chapter, we actually show that we can view the noisy quantum theory *as a special case of the noiseless quantum theory*. This relation may seem bizarre at first, but the purification theorem allows us to make this connection. The quantum theory that we present in this chapter is a noiseless quantum theory, but we name it *the purified quantum theory*, in order to distinguish it from the description of the noiseless quantum theory in Chapter 3.

The purified quantum theory shows that it is possible to view noise as resulting from entanglement of a system with another system. We have actually seen a glimpse of this phenomenon in the previous chapter when we introduced the notion of the local density operator, but we did not highlight it in detail there. The example was the maximally entangled Bell state $|\Phi^+\rangle^{AB}$. This state is a pure state on the two systems A and B, but the local density operator of Alice is the maximally mixed state π^A. We saw that the local

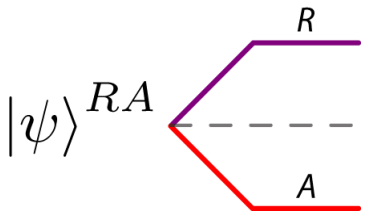

Figure 5.1: The above diagram denotes the purification $|\psi\rangle^{RA}$ of a density matrix ρ^A. The above diagram indicates that the reference system R is entangled with the system A. The purification theorem states that the noise inherent in a density matrix ρ^A is due to entanglement with a reference system R. We typically use the color purple throughout to indicate a system that is not accessible to the parties involved in a protocol.

density operator is a mathematical object that allows us to make all the predictions about any local measurement or evolution. We also have seen that a density operator arises from an ensemble, but there is also the reverse interpretation, that an ensemble corresponds to the spectral decomposition of any density operator. There is a sense in which we can view this local density operator as arising from an ensemble where we choose the states $|0\rangle$ and $|1\rangle$ with equal probability 1/2. The purification idea goes as far to say that the noisy ensemble for Alice with density operator π^A arises from the entanglement of her system with Bob's. We explore this idea in more detail in this final chapter on the quantum theory.

5.1 Purification

Suppose we are given a density operator ρ^A on a system A and suppose that its spectral decomposition is as follows:

$$\rho^A = \sum_x p_X(x)|x\rangle\langle x|^A. \tag{5.1}$$

We can associate the ensemble $\{p_X(x), |x\rangle\}$ to this density operator according to its spectral decomposition.

Definition 5.1.1 (Purification)**.** *A purification of ρ^A is a pure bipartite state $|\psi\rangle^{RA}$ on a reference system R and the original system A. The purification state $|\psi\rangle^{RA}$ has the property that the reduced state on system A is equal to ρ^A in (5.1):*

$$\rho^A = Tr_R\left\{|\psi\rangle\langle\psi|^{RA}\right\}. \tag{5.2}$$

Any density operator ρ^A has a purification $|\psi\rangle^{RA}$. We claim that the following state $|\psi\rangle^{RA}$ is a purification of ρ^A:

$$|\psi\rangle^{RA} \equiv \sum_x \sqrt{p_X(x)}|x\rangle^R|x\rangle^A, \tag{5.3}$$

where the set $\{|x\rangle^R\}_x$ of vectors are some set of orthonormal vectors for the reference system R. The next exercise asks you to verify this claim.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 5.1.1 Show that the state $|\psi\rangle^{RA}$, as defined in the above proof, is a purification of the density operator ρ^A.

The purification idea has an interesting physical implication—it implies that we can think of our lack of knowledge about a particular quantum system as being due to entanglement with some external reference system to which we do not have access. That is, we can think that the density operator ρ^A with corresponding ensemble $\{p_X(x), |x\rangle\}$ arises from the entanglement of the system A with the reference system R and from our lack of access to the system R.

Stated another way, the purification idea gives us a fundamentally different way to interpret noise. The interpretation is that any noise on a local system is due to entanglement with another system to which we do not have access. This interpretation extends to the noise from a noisy quantum channel. We can view this noise as arising from the interaction of the system that we possess with an external environment over which we have no control.

The global state is a pure state, but a reduced state is not a pure state in general because we trace over the reference system to obtain it. A reduced state is pure if and only if the global state is a pure product state.

Exercise 5.1.2 Show that all purifications are related by a unitary operator on the reference system.

Exercise 5.1.3 Find a purification of the following classical-quantum state:

$$\sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A. \tag{5.4}$$

Exercise 5.1.4 Let $\{p_X(x), \rho_x^A\}$ be an ensemble of density operators. Suppose that $|\psi_x\rangle^{RA}$ is a purification of ρ_x^A. The expected density operator of the ensemble is

$$\rho^A \equiv \sum_x p_X(x)\rho_x^A. \tag{5.5}$$

Find a purification of ρ^A.

5.1.1 Extension of a Quantum System

We can also define an *extension* of a quantum system ρ^A. It is some noisy quantum system Ω^{RA} such that

$$\rho^A = \mathrm{Tr}_R\{\Omega^{RA}\}. \tag{5.6}$$

This definition is useful sometimes, but we can always find a purification of the extended state.

5.2 Isometric Evolution

A noisy quantum channel admits a purification as well. We motivate this idea with a simple example.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

5.2.1 Isometric Extension of the Bit-Flip Channel

Consider the bit-flip channel from (4.159)—it applies the identity operator with some probability $1-p$ and applies the bit flip Pauli operator X with probability p. Suppose that we input a qubit system A in the state $|\psi\rangle$ to this channel. The ensemble corresponding to the state at the output has the following form:

$$\{\{1-p, |\psi\rangle\}, \{p, X|\psi\rangle\}\}, \tag{5.7}$$

and the density operator of the resulting state is

$$(1-p)|\psi\rangle\langle\psi| + pX|\psi\rangle\langle\psi|X. \tag{5.8}$$

The following state is a purification of the above density operator (you should quickly check that this relation holds):

$$\sqrt{1-p}|\psi\rangle^A|0\rangle^E + \sqrt{p}X|\psi\rangle^A|1\rangle^E. \tag{5.9}$$

We label the original system as A and label the purification system as E. In this context, we can view the purification system as the environment of the channel.

There is another way for interpreting the dynamics of the above bit-flip channel. Instead of determining the ensemble for the channel and then purifying, we can say that the channel directly implements the following map from the system A to the larger joint system AE:

$$|\psi\rangle^A \rightarrow \sqrt{1-p}|\psi\rangle^A|0\rangle^E + \sqrt{p}X|\psi\rangle^A|1\rangle^E. \tag{5.10}$$

We see that any positive p, i.e., any amount of noise in the channel, can lead to entanglement of the input system with the environment E. We then obtain the noisy dynamics of the channel by discarding (tracing out) the environment system E.

Exercise 5.2.1 Find two input states for which the map in (5.10) does not lead to entanglement between systems A and E.

The above map is an *isometric extension* of the bit-flip channel. Let us label it as $U^{A\rightarrow AE}$ where the notation indicates that the input system is A and the output system is AE. An isometry is similar to a unitary operator but different because it maps states on one input system to states on a joint system. It does not admit a square matrix representation, but instead admits a rectangular matrix representation. The matrix representation of this isometric operation consists of the following matrix elements:

$$\begin{bmatrix} \langle 0|^A\langle 0|^E U^{A\rightarrow AE}|0\rangle^A & \langle 0|^A\langle 0|^E U^{A\rightarrow AE}|1\rangle^A \\ \langle 0|^A\langle 1|^E U^{A\rightarrow AE}|0\rangle^A & \langle 0|^A\langle 1|^E U^{A\rightarrow AE}|1\rangle^A \\ \langle 1|^A\langle 0|^E U^{A\rightarrow AE}|0\rangle^A & \langle 1|^A\langle 0|^E U^{A\rightarrow AE}|1\rangle^A \\ \langle 1|^A\langle 1|^E U^{A\rightarrow AE}|0\rangle^A & \langle 1|^A\langle 1|^E U^{A\rightarrow AE}|1\rangle^A \end{bmatrix} = \begin{bmatrix} \sqrt{1-p} & 0 \\ 0 & \sqrt{p} \\ 0 & \sqrt{1-p} \\ \sqrt{p} & 0 \end{bmatrix}. \tag{5.11}$$

There is no reason that we have to choose the environment states as we did in (5.10). We could have chosen the environment states to be any orthonormal basis—isometric behavior only requires that the states on the environment be distinguishable. This is related to the fact that all purifications are related by a unitary on the purifying system (see Exercise 5.1.2).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

An Isometry is Part of a Unitary on a Larger System

We can view the dynamics in (5.10) as an interaction between an initially pure environment and the qubit state $|\psi\rangle$. So, an equivalent way to implement the isometric mapping is with a two-step procedure. We first assume that the environment of the channel is in a pure state $|0\rangle^E$ before the interaction begins. The joint state of the qubit $|\psi\rangle$ and the environment is

$$|\psi\rangle^A|0\rangle^E. \tag{5.12}$$

These two systems then interact according to a unitary operator. We can specify two columns of the unitary operator (we make this more clear in a bit) by means of the isometric mapping in (5.10):

$$V^{AE}|\psi\rangle^A|0\rangle^E = \sqrt{1-p}|\psi\rangle^A|0\rangle^E + \sqrt{p}X|\psi\rangle^A|1\rangle^E. \tag{5.13}$$

In order to specify the full unitary V^{AE}, we must also specify how it behaves when the initial state of the qubit and the environment is

$$|\psi\rangle^A|1\rangle^E. \tag{5.14}$$

We choose the mapping to be as follows so that the overall interaction is unitary:

$$V^{AE}|\psi\rangle^A|1\rangle^E = \sqrt{p}|\psi\rangle^A|0\rangle^E - \sqrt{1-p}X|\psi\rangle^A|1\rangle^E. \tag{5.15}$$

Exercise 5.2.2 Check that the operator V^{AE} is unitary by determining its action on the computational basis $\left\{|0\rangle^A|0\rangle^E, |0\rangle^A|1\rangle^E, |1\rangle^A|0\rangle^E, |1\rangle^A|1\rangle^E\right\}$ and showing that all of the outputs for each of these inputs forms an orthonormal basis.

Exercise 5.2.3 Verify that the matrix representation of the full unitary operator V^{AE} is

$$\begin{bmatrix} \sqrt{1-p} & \sqrt{p} & 0 & 0 \\ 0 & 0 & \sqrt{p} & -\sqrt{1-p} \\ 0 & 0 & \sqrt{1-p} & \sqrt{p} \\ \sqrt{p} & -\sqrt{1-p} & 0 & 0 \end{bmatrix}, \tag{5.16}$$

by considering the matrix elements $\langle i|^A\langle j|^E V|k\rangle^A|l\rangle^E$.

The Complementary Channel

We may not only be interested in the receiver's output of the quantum channel. We may also be interested in determining the environment's output from the channel. This idea becomes increasingly important as we proceed in our study of quantum Shannon theory. We should consider all parties in a quantum protocol, and the purified quantum theory allows us to do so. We consider the environment as one of the parties in a quantum protocol because the environment could also be receiving some quantum information from the sender.

We can obtain the environment's output from the quantum channel simply by tracing out every system besides the environment. The map from the sender to the environment is

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

known as a *complementary channel*. In our example of the isometric extension of the bit flip channel in (5.10), we can check that the environment receives the following density operator

$$\begin{aligned}
&\text{Tr}_A\Big\{\Big(\sqrt{1-p}|\psi\rangle^A|0\rangle^E + \sqrt{p}X|\psi\rangle^A|1\rangle^E\Big)\Big(\sqrt{1-p}\langle\psi|^A\langle 0|^E + \sqrt{p}\langle\psi|^A X\langle 1|^E\Big)\Big\} \\
&= \text{Tr}_A\Big\{(1-p)|\psi\rangle\langle\psi|^A|0\rangle\langle 0|^E + \sqrt{p(1-p)}X|\psi\rangle\langle\psi|^A|1\rangle\langle 0|^E\Big\} \\
&\qquad + \text{Tr}_A\Big\{\sqrt{p(1-p)}|\psi\rangle\langle\psi|^A X|0\rangle\langle 1|^E + pX|\psi\rangle\langle\psi|^A X|1\rangle\langle 1|^E\Big\} \qquad (5.17) \\
&= (1-p)|0\rangle\langle 0|^E + \sqrt{p(1-p)}\langle\psi|X|\psi\rangle|1\rangle\langle 0|^E \\
&\qquad + \sqrt{p(1-p)}\langle\psi|X|\psi\rangle|0\rangle\langle 1|^E + p|1\rangle\langle 1|^E \qquad (5.18) \\
&= (1-p)|0\rangle\langle 0|^E + \sqrt{p(1-p)}\langle\psi|X|\psi\rangle\Big(|1\rangle\langle 0|^E + |0\rangle\langle 1|^E\Big) + p|1\rangle\langle 1|^E \qquad (5.19) \\
&= (1-p)|0\rangle\langle 0|^E + \sqrt{p(1-p)}2\,\text{Re}\{\alpha^*\beta\}\Big(|1\rangle\langle 0|^E + |0\rangle\langle 1|^E\Big) + p|1\rangle\langle 1|^E, \qquad (5.20)
\end{aligned}$$

where in the last line we assume that the qubit $|\psi\rangle \equiv \alpha|0\rangle + \beta|1\rangle$.

It is helpful to examine several cases of the above example. Consider the case where the noise parameter $p = 0$ or $p = 1$. In this case, the environment receives one of the respective states $|0\rangle$ or $|1\rangle$. Therefore, in these cases, the environment does not receive any of the quantum information about the state $|\psi\rangle$ transmitted down the channel—it does not learn anything about the probability amplitudes α or β. This viewpoint is a completely different way to see that the channel is truly noiseless in these cases. A channel is noiseless if the environment of the channel does not learn anything about the states that we transmit through it, i.e, the channel does not leak quantum information to the environment. Now let us consider the case where $0 < p < 1$. As p approaches $1/2$ from either above or below, the amplitude $\sqrt{p(1-p)}$ of the off-diagonal terms is a monotonic function that reaches its peak at $1/2$. Thus, at the peak $1/2$, the off-diagonal terms are the strongest, implying that the environment is stealing much of the coherence from the original quantum state $|\psi\rangle$.

Exercise 5.2.4 Show that the receiver's output density operator for a bit-flip channel with $p = 1/2$ is the same as what the environment obtains.

5.2.2 Isometric Extension of a General Noisy Quantum Channel

We now discuss the general definition of an isometric extension of a quantum channel. Let $\mathcal{N}^{A\to B}$ denote a noisy quantum channel, where the notation indicates that its input is some quantum system A and its output is some quantum system B. The *isometric extension* or *Stinespring dilation* of a quantum channel is the purification of that channel. The isometric extension $U_{\mathcal{N}}^{A\to BE}$ of quantum channel $\mathcal{N}^{A\to B}$ is a particular map related to $\mathcal{N}^{A\to B}$. The input to the isometry $U_{\mathcal{N}}^{A\to BE}$ is the original input system A, and the output of the isometry is the channel output B and an environment system E (the environment system is analogous to the purification system from Section 5.1). The notation $U_{\mathcal{N}}^{A\to BE}$ indicates the input and output systems, and Figure 5.2 depicts a quantum circuit for the isometric extension. An isometry possesses the following two properties:

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

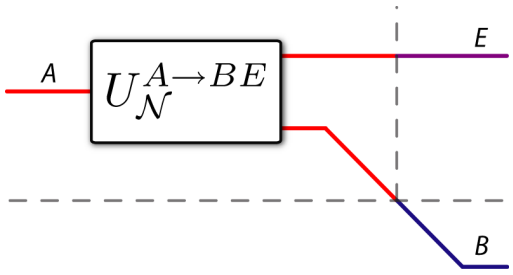

Figure 5.2: The above figure depicts the isometric extension $U_{\mathcal{N}}^{A\to BE}$ of a quantum channel $\mathcal{N}^{A\to B}$. The extension $U_{\mathcal{N}}^{A\to BE}$ includes the inaccessible enviroment on system E as a "receiver" of quantum information. Ignoring the environment E gives the noisy channel $\mathcal{N}^{A\to B}$.

1. It produces the evolution of the noisy quantum channel $\mathcal{N}^{A\to B}$ if we trace out the environment system:
$$\text{Tr}_E\{U_{\mathcal{N}}^{A\to BE}(\rho)\} = \mathcal{N}^{A\to B}(\rho), \tag{5.21}$$
where ρ is any density operator input to the channel $\mathcal{N}^{A\to B}$.

2. It behaves as an *isometry*—it is analogous to a rectangular matrix that behaves somewhat like a unitary operator. The matrix representation of an isometry is a rectangular matrix formed from selecting only a few of the columns from a unitary matrix. An isometry obeys the following two properties:
$$U_{\mathcal{N}}^{\dagger}U_{\mathcal{N}} = I^A, \tag{5.22}$$
$$U_{\mathcal{N}}U_{\mathcal{N}}^{\dagger} = \Pi^{BE}, \tag{5.23}$$
where Π^{BE} is some projector on the joint system BE. The first property indicates that the isometry behaves analogously to a unitary operator, because we can determine an inverse operation simply by taking its conjugate transpose. The name "isometry" derives from the first property because it implies that the mapping preserves the lengths of vectors. The second property distinguishes an isometric operation from a unitary one. It states that the isometry takes states in the input system A to a particular subspace of the joint system BE. The projector Π^{BE} projects onto the subspace where the isometry takes input quantum states.

Isometric Extension from the Kraus Operators

It is possible to determine the isometric extension of a quantum channel directly from its Kraus operators. Consider a noisy quantum channel $\mathcal{N}^{A\to B}$ with the following Kraus representation:
$$\mathcal{N}^{A\to B}(\rho^A) = \sum_j N_j \rho^A N_j^{\dagger}. \tag{5.24}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The isometric extension of the channel $\mathcal{N}^{A\to B}$ is the following map:

$$U_{\mathcal{N}}^{A\to BE} \equiv \sum_j N_j \otimes |j\rangle^E. \tag{5.25}$$

It is straightforward to verify that the above map is an isometry. We first need to verify that $U_{\mathcal{N}}^\dagger U_{\mathcal{N}}$ is equal to the identity on the system A:

$$\begin{aligned} U_{\mathcal{N}}^\dagger U_{\mathcal{N}} &= \left(\sum_k N_k^\dagger \otimes \langle k|^E\right)\left(\sum_j N_j \otimes |j\rangle^E\right) && (5.26)\\ &= \sum_{k,j} N_k^\dagger N_j \langle k|j\rangle && (5.27)\\ &= \sum_k N_k^\dagger N_k && (5.28)\\ &= I. && (5.29)\end{aligned}$$

The last equality follows from the completeness condition of the Kraus operators. We next need to prove that $U_{\mathcal{N}} U_{\mathcal{N}}^\dagger$ is a projector on the joint system BE. This follows simply by noting that

$$U_{\mathcal{N}} U_{\mathcal{N}}^\dagger U_{\mathcal{N}} U_{\mathcal{N}}^\dagger = U_{\mathcal{N}}\Big(U_{\mathcal{N}}^\dagger U_{\mathcal{N}}\Big)U_{\mathcal{N}}^\dagger = U_{\mathcal{N}}\ I^A\ U_{\mathcal{N}}^\dagger = U_{\mathcal{N}} U_{\mathcal{N}}^\dagger. \tag{5.30}$$

We can also prove this a bit more explicitly. First, consider that

$$\begin{aligned} U_{\mathcal{N}} U_{\mathcal{N}}^\dagger &= \left(\sum_j N_j \otimes |j\rangle^E\right)\left(\sum_k N_k^\dagger \otimes \langle k|^E\right) && (5.31)\\ &= \sum_{j,k} N_j N_k^\dagger \otimes |j\rangle\langle k|^E && (5.32)\end{aligned}$$

Let us now verify that the above operator is a projector by squaring it:

$$\begin{aligned} \left(U_{\mathcal{N}} U_{\mathcal{N}}^\dagger\right)^2 &= \left(\sum_{j,k} N_j N_k^\dagger \otimes |j\rangle\langle k|^E\right)\left(\sum_{j',k'} N_{j'} N_{k'}^\dagger \otimes |j'\rangle\langle k'|^E\right) && (5.33)\\ &= \sum_{j,k,j',k'} N_j N_k^\dagger N_{j'} N_{k'}^\dagger \otimes |j\rangle\langle k|j'\rangle\langle k'|^E && (5.34)\\ &= \sum_{j,k,k'} N_j N_k^\dagger N_k N_{k'}^\dagger \otimes |j\rangle\langle k'|^E && (5.35)\\ &= \sum_{j,k'} N_j \left(\sum_k N_k^\dagger N_k\right) N_{k'}^\dagger \otimes |j\rangle\langle k'|^E && (5.36)\\ &= \sum_{j,k'} N_j N_{k'}^\dagger \otimes |j\rangle\langle k'|^E && (5.37)\\ &= U_{\mathcal{N}} U_{\mathcal{N}}^\dagger. && (5.38)\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Finally, we should verify that $U_{\mathcal{N}}$ is an extension of $\mathcal{N}$. The application of the isometry to an arbitrary density operator ρ^A gives the following map:

$$U_{\mathcal{N}}^{A\to BE}(\rho^A) \equiv U_{\mathcal{N}}\rho^A U_{\mathcal{N}}^\dagger \tag{5.39}$$

$$= \left(\sum_j N_j \otimes |j\rangle^E\right)\rho^A\left(\sum_k N_k^\dagger \otimes \langle k|^E\right) \tag{5.40}$$

$$= \sum_{j,k} N_j \rho^A N_k^\dagger \otimes |j\rangle\langle k|^E, \tag{5.41}$$

and tracing out the environment system gives back the original noisy channel $\mathcal{N}^{A\to B}$:

$$\mathrm{Tr}_E\{U_{\mathcal{N}}^{A\to BE}(\rho^A)\} = \sum_j N_j \rho^A N_j^\dagger \tag{5.42}$$

$$= \mathcal{N}^{A\to B}(\rho^A). \tag{5.43}$$

Exercise 5.2.5 Show that all isometric extensions of a noisy quantum channel are equivalent up to an isometry on the environment system (this is similar to the result of Exercise 5.1.2).

Exercise 5.2.6 Show that the isometric extension of the erasure channel is

$$U_{\mathcal{N}}^{A\to BE} = \sqrt{1-\varepsilon}(|0\rangle^B\langle 0|^A + |1\rangle^B\langle 1|^A)\otimes|e\rangle^E + \sqrt{\varepsilon}|e\rangle^B\langle 0|^A\otimes|0\rangle^E + \sqrt{\varepsilon}|e\rangle^B\langle 1|^A\otimes|1\rangle^E. \tag{5.44}$$

Exercise 5.2.7 Determine the resulting state when Alice inputs an arbitrary pure state $|\psi\rangle$ into the erasure channel. Verify that Bob and Eve receive the same ensemble (they have the same local density operator) when the erasure probability $\varepsilon = 1/2$.

Exercise 5.2.8 Show that the matrix representation of the isometric extension of the erasure channel is

$$\begin{bmatrix} \langle 0|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 0|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle 0|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 0|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle 0|^B\langle e|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 0|^B\langle e|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle 1|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 1|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle 1|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 1|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle 1|^B\langle e|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 1|^B\langle e|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle e|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle e|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle e|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle e|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle e|^B\langle e|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle e|^B\langle e|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \sqrt{1-\varepsilon} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & \sqrt{1-\varepsilon} \\ \sqrt{\varepsilon} & 0 \\ 0 & \sqrt{\varepsilon} \\ 0 & 0 \end{bmatrix}. \tag{5.45}$$

Exercise 5.2.9 Show that the matrix representation of the isometric extension $U_{\mathcal{N}}^{A\to BE}$ of the amplitude damping channel is

$$\begin{bmatrix} \langle 0|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 0|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle 0|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 0|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle 1|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 1|^B\langle 0|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \\ \langle 1|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|0\rangle^A & \langle 1|^B\langle 1|^E U_{\mathcal{N}}^{A\to BE}|1\rangle^A \end{bmatrix} = \begin{bmatrix} 0 & \sqrt{\gamma} \\ 1 & 0 \\ 0 & 0 \\ 0 & \sqrt{1-\gamma} \end{bmatrix}. \tag{5.46}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 5.2.10 Consider a full unitary $V^{AE\to BE}$ such that

$$Tr_E\Big\{V\Big(\rho^A\otimes|0\rangle\langle 0|^E\Big)V^\dagger\Big\} \tag{5.47}$$

gives the amplitude damping channel. Show that the matrix representation of V is

$$\begin{bmatrix} \langle 0|^B\langle 0|^E V|0\rangle^A|0\rangle^E & \langle 0|^B\langle 0|^E V|0\rangle^A|1\rangle^E & \langle 0|^B\langle 0|^E V|1\rangle^A|0\rangle^E & \langle 0|^B\langle 0|^E V|1\rangle^A|1\rangle^E \\ \langle 0|^B\langle 1|^E V|0\rangle^A|0\rangle^E & \langle 0|^B\langle 1|^E V|0\rangle^A|1\rangle^E & \langle 0|^B\langle 1|^E V|1\rangle^A|0\rangle^E & \langle 0|^B\langle 1|^E V|1\rangle^A|1\rangle^E \\ \langle 1|^B\langle 0|^E V|0\rangle^A|0\rangle^E & \langle 1|^B\langle 0|^E V|0\rangle^A|1\rangle^E & \langle 1|^B\langle 0|^E V|1\rangle^A|0\rangle^E & \langle 1|^B\langle 0|^E V|1\rangle^A|1\rangle^E \\ \langle 1|^B\langle 1|^E V|0\rangle^A|0\rangle^E & \langle 1|^B\langle 1|^E V|0\rangle^A|1\rangle^E & \langle 1|^B\langle 1|^E V|1\rangle^A|0\rangle^E & \langle 1|^B\langle 1|^E V|1\rangle^A|1\rangle^E \end{bmatrix}$$
$$= \begin{bmatrix} 0 & -\sqrt{1-\gamma} & \sqrt{\gamma} & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \sqrt{\gamma} & \sqrt{1-\gamma} & 0 \end{bmatrix}. \tag{5.48}$$

Exercise 5.2.11 Consider the full unitary operator for the amplitude damping channel from the previous exercise. Show that the density operator

$$Tr_B\Big\{V\Big(\rho^A\otimes|0\rangle\langle 0|^E\Big)V^\dagger\Big\} \tag{5.49}$$

that Eve receives has the following matrix representation:

$$\begin{bmatrix} \gamma p & \sqrt{\gamma}\eta^* \\ \sqrt{\gamma}\eta & 1-\gamma p \end{bmatrix} \tag{5.50}$$

if

$$\rho^A = \begin{bmatrix} 1-p & \eta \\ \eta^* & p \end{bmatrix}. \tag{5.51}$$

By comparing with (4.231), observe that the output to Eve is the bit flip of the output of an amplitude damping channel with damping parameter $1-\gamma$.

Exercise 5.2.12 Consider the amplitude damping channel with parameter γ. Show that it is possible to simulate the channel to Eve by first sending Alice's density operator through an amplitude damping channel with parameter $1-\gamma$ and swapping A with E, i.e., show that

$$\mathrm{Tr}_B\Big\{V_\gamma\Big(\rho^A\otimes|0\rangle\langle 0|^E\Big)V_\gamma^\dagger\Big\} = \mathrm{Tr}_E\Big\{SV_{1-\gamma}\Big(\rho^A\otimes|0\rangle\langle 0|^E\Big)V_{1-\gamma}^\dagger S^\dagger\Big\}, \tag{5.52}$$

where V_γ is the unitary operator for an amplitude damping channel with parameter γ and S is the swapping operator.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Complementary channel

In the purified quantum theory, it is useful to consider all parties that are participating in a given protocol. One such party is the environment of the channel and we call her Eve, because she corresponds to an eavesdropper in the cryptographic setting.

For any quantum channel $\mathcal{N}^{A\to B}$, there exists an isometric extension $U_{\mathcal{N}}^{A\to BE}$ of that channel. The complementary channel $(\mathcal{N}^c)^{A\to E}$ is a quantum channel from Alice to Eve. We obtain it by tracing out Bob's system from the output of the isometry:

$$(\mathcal{N}^c)^{A\to E}(\rho) \equiv \operatorname{Tr}_B\{U_{\mathcal{N}}^{A\to BE}(\rho)\}. \tag{5.53}$$

It captures the noise that Eve "sees" by having her system coupled to Bob's system.

Exercise 5.2.13 Show that Eve's density operator is of the following form:

$$\rho \to \sum_{i,j} \operatorname{Tr}\left\{N_i \rho N_j^\dagger\right\}|i\rangle\langle j|, \tag{5.54}$$

if we take the isometric extension of the channel to be of the form in (5.25).

The complementary channel is unique only up to an isometry on Eve's system. It inherits this property from the fact that an isometric extension of a noisy channel is unique only up to isometries on Eve's system. For all practical purposes, this lack of uniqueness does not affect our study of the noise that Eve sees because the measures of noise in Chapter 11 are invariant under isometries on Eve's system.

5.2.3 Generalized Dephasing Channels

A generalized dephasing channel is one that preserves states diagonal in some preferred orthonormal basis $\{|x\rangle\}$, but it can add arbitrary phases to the off-diagonal elements of a density operator in this basis. The isometry of a generalized dephasing channel acts as follows on the basis $\{|x\rangle\}$:

$$U_{\mathcal{N}_{\mathrm{D}}}^{A'\to BE}|x\rangle^{A'} = |x\rangle^B|\varphi_x\rangle^E, \tag{5.55}$$

where $|\varphi_x\rangle^E$ is some state for the environment (these states need not be mutually orthogonal). Thus, we can represent the isometry as follows:

$$U_{\mathcal{N}_{\mathrm{D}}}^{A'\to BE} \equiv \sum_x |x\rangle^B|\varphi_x\rangle^E\langle x|^{A'}, \tag{5.56}$$

and its action on a density operator ρ is

$$U_{\mathcal{N}_{\mathrm{D}}}\rho U_{\mathcal{N}_{\mathrm{D}}}^\dagger = \sum_{x,x'}\langle x|\rho|x'\rangle \ \ |x\rangle\langle x'|^B \otimes |\varphi_x\rangle\langle\varphi_{x'}|^E. \tag{5.57}$$

Tracing out the environment gives the action of the channel $\mathcal{N}_{\mathrm{D}}$ to the receiver

$$\mathcal{N}_{\mathrm{D}}(\rho) = \sum_{x,x'}\langle x|\rho|x'\rangle\langle\varphi_{x'}|\varphi_x\rangle \ \ |x\rangle\langle x'|^B, \tag{5.58}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where we observe that this channel preserves the diagonal components $\{|x\rangle\langle x|\}$ of ρ, but it multiplies the $d(d-1)$ off-diagonal elements of ρ by arbitrary phases, depending on the $d(d-1)$ overlaps $\langle\varphi_{x'}|\varphi_x\rangle$ of the environment states (where $x \neq x'$). Tracing out the receiver gives the action of the complementary channel $\mathcal{N}_{\mathrm{D}}^{c}$ to the environment

$$\mathcal{N}_{\mathrm{D}}^{c}(\rho) = \sum_x \langle x|\rho|x\rangle \ |\varphi_x\rangle\langle\varphi_x|^E. \tag{5.59}$$

Observe that the channel to the environment is entanglement-breaking. That is, the action of the channel is the same as first performing a von Neumann measurement in the basis $\{|x\rangle\}$ and preparing a state $|\varphi_x\rangle^E$ conditional on the outcome of the measurement (it is a classical-quantum channel from Section 4.4.6). Additionally, the receiver Bob can simulate the action of this channel to the receiver by performing the same actions on the state that he receives.

Exercise 5.2.14 Explicitly show that the following qubit dephasing channel is a special case of a generalized dephasing channel:

$$\rho \to (1-p)\rho + pZ\rho Z. \tag{5.60}$$

5.2.4 Quantum Hadamard Channels

Quantum Hadamard channels are those whose complements are entanglement-breaking. We can write its output as the Hadamard product (element-wise multiplication) of a representation of the input density operator with another operator. To discuss how this comes about, suppose that the complementary channel $(\mathcal{N}^c)^{A'\to E}$ of a channel $\mathcal{N}^{A'\to B}$ is entanglement-breaking. Then, using the fact that its Kraus operators $|\xi_i\rangle^E\langle\zeta_i|^{A'}$ are unit rank (see Theorem 4.4.1) and the construction in (5.25) for an isometric extension, we can write an isometric extension $U_{\mathcal{N}^c}$ for $\mathcal{N}^c$ as

$$U_{\mathcal{N}^c}\rho U_{\mathcal{N}^c}^\dagger = \sum_{i,j} |\xi_i\rangle^E\langle\zeta_i|^{A'}\rho|\zeta_j\rangle^{A'}\langle\xi_j|^E \otimes |i\rangle^B\langle j|^B \tag{5.61}$$

$$= \sum_{i,j} \langle\zeta_i|^{A'}\rho|\zeta_j\rangle^{A'}|\xi_i\rangle^E\langle\xi_j|^E \otimes |i\rangle^B\langle j|^B. \tag{5.62}$$

The sets $\{|\xi_i\rangle^E\}$ and $\{|\zeta_i\rangle^{A'}\}$ each do not necessarily consist of orthonormal states, but the set $\{|i\rangle^B\}$ does because it is the environment of the complementary channel. Tracing over the system E gives the original channel from system A' to B:

$$\mathcal{N}_{\mathrm{H}}^{A'\to B}(\rho) = \sum_{i,j} \langle\zeta_i|^{A'}\rho|\zeta_j\rangle^{A'}\langle\xi_j|\xi_i\rangle^E|i\rangle^B\langle j|^B. \tag{5.63}$$

Let Σ denote the matrix with elements $[\Sigma]_{i,j} = \langle\zeta_i|^{A'}\rho|\zeta_j\rangle^{A'}$, a representation of the input state ρ, and let Γ denote the matrix with elements $[\Gamma]_{i,j} = \langle\xi_i|\xi_j\rangle^E$. Then, from (5.63), it is

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

clear that the output of the channel is the Hadamard product $*$ of Σ and $\Gamma^\dagger$ with respect to the basis $\{|i\rangle^B\}$:

$$\mathcal{N}_{\mathrm{H}}^{A'\to B}(\rho) = \Sigma * \Gamma^\dagger. \tag{5.64}$$

For this reason, such a channel is known as a Hadamard channel.

Hadamard channels are degradable, meaning that there exists a degrading map $\mathcal{D}^{B\to E}$ such that Bob can simulate the channel to Eve:

$$\forall \rho \quad \mathcal{D}^{B\to E}(\mathcal{N}_{\mathrm{H}}^{A'\to B}(\rho)) = (\mathcal{N}_{\mathrm{H}}^{c})^{A'\to E}(\rho). \tag{5.65}$$

If Bob performs a von Neumann measurement of his state in the basis $\{|i\rangle^B\}$ and prepares the state $|\xi_i\rangle^E$ conditional on the outcome of the measurement, this procedure simulates the complementary channel $(\mathcal{N}^c)^{A'\to E}$ and also implies that the degrading map $\mathcal{D}^{B\to E}$ is entanglement-breaking. To be more precise, the Kraus operators of the degrading map $\mathcal{D}^{B\to E}$ are $\{|\xi_i\rangle^E\langle i|^B\}$ so that

$$\mathcal{D}^{B\to E}(\mathcal{N}_{\mathrm{H}}^{A'\to B}(\sigma)) = \sum_i |\xi_i\rangle^E\langle i|^B \mathcal{N}^{A'\to B}(\sigma)|i\rangle^B\langle\xi_i|^E \tag{5.66}$$

$$= \sum_i \langle i|^{A'}\sigma|i\rangle^{A'}|\xi_i\rangle\langle\xi_i|^E, \tag{5.67}$$

demonstrating that this degrading map effectively simulates the complementary channel $(\mathcal{N}_{\mathrm{H}}^{c})^{A'\to E}$. Note that we can view this degrading map as the composition of two maps: a first map $\mathcal{D}_1^{B\to Y}$ performs the von Neumann measurement, leading to a classical variable Y, and a second map $\mathcal{D}_2^{Y\to E}$ performs the state preparation, conditional on the value of the classical variable Y. We can therefore write $\mathcal{D}^{B\to E} = \mathcal{D}_2^{Y\to E} \circ \mathcal{D}_1^{B\to Y}$. This particular form of the channel has implications for its quantum capacity (see Chapter 23) and its more general capacities (see Chapter 24). Observe that a generalized dephasing channel from the previous section is a quantum Hadamard channel because the map to its environment is entanglement-breaking.

5.3 Coherent Quantum Instrument

It is useful to consider the isometric extension of a quantum instrument (we discussed quantum instruments in Section 4.4.7). This viewpoint is important when we recall that a quantum instrument is the most general map from a quantum system to a quantum system and a classical system.

Recall from Section 4.4.7 that a quantum instrument acts as follows:

$$\rho \to \sum_j \mathcal{E}_j^{A\to B}(\rho) \otimes |j\rangle\langle j|^J, \tag{5.68}$$

where each $\mathcal{E}_j$ is a completely-positive trade-reducing (CPTR) map that has the following form:

$$\mathcal{E}_j^{A\to B}(\rho) = \sum_k M_{j,k}\rho M_{j,k}^\dagger, \tag{5.69}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and the operators $M_{j,k}$ satisfy the relation $\sum_k M_{j,k}^\dagger M_{j,k} \leq I$.

We now describe a particular coherent evolution that implements the above transformation when we trace over certain degrees of freedom. An isometric extension of each CPTR map $\mathcal{E}_j$ is as follows:

$$U_{\mathcal{E}_j}^{A\to BE} \equiv \sum_k M_{j,k} \otimes |k\rangle^E, \tag{5.70}$$

where the operator $M_{j,k}$ acts on the input state and the environment system E is large enough to accomodate all of the CPTR maps $\mathcal{E}_j$. That is, if the first map $\mathcal{E}_1$ has states $\left\{|1\rangle^E, \ldots, |d_1\rangle^E\right\}$, then the second map $\mathcal{E}_2$ has states $\left\{|d_1+1\rangle^E, \ldots, |d_1+d_2\rangle^E\right\}$ so that the states on E are orthogonal for all the different maps $\mathcal{E}_j$ that are part of the instrument. We can embed this isometric extension into the evolution in (5.68) as follows:

$$\rho \to \sum_j U_{\mathcal{E}_j}^{A\to BE}(\rho) \otimes |j\rangle\langle j|^J. \tag{5.71}$$

This evolution is not quite fully coherent, but a simple modification of it does make it fully coherent:

$$\sum_j U_{\mathcal{E}_j}^{A\to BE} \otimes |j\rangle^J \otimes |j\rangle^{E_J}. \tag{5.72}$$

The full action of the coherent instrument is then as follows:

$$\rho \to \sum_j U_{\mathcal{E}_j}^{A\to BE}(\rho) \otimes |j\rangle\langle j'|^J \otimes |j\rangle\langle j'|^{E_J} \tag{5.73}$$

$$= \sum_{j,k,j',k'} M_{j,k}\rho M_{j',k'}^\dagger \otimes |k\rangle\langle k'|^E \otimes |j\rangle\langle j'|^J \otimes |j\rangle\langle j'|^{E_J}. \tag{5.74}$$

One can then check that tracing over the environmental degrees of freedom E and E_J reproduces the action of the quantum instrument in (5.68).

5.4 Coherent Measurement

We end this chapter by discussing a coherent measurement. This last section shows that it is sufficient to describe all of the quantum theory in the so-called "traditionalist" way by using only unitary evolutions and von Neumann projective measurements.

Suppose that we have a set of measurement operators $\{M_j\}_j$ such that $\sum_j M_j^\dagger M_j = I$. In the noisy quantum theory, we found that the post-measurement state of a measurement on a quantum system S with density operator ρ is

$$\frac{M_j \rho M_j^\dagger}{p_J(j)}, \tag{5.75}$$

where the measurement outcome j occurs with probability

$$p_J(j) = \text{Tr}\left\{M_j^\dagger M_j \rho\right\}. \tag{5.76}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We would like a way to perform the above measurement on system S in a *coherent* fashion. The isometry in (5.25) gives a hint for how we can structure such a coherent measurement. We can build the coherent measurement as the following isometry:

$$U^{S\to SS'} \equiv \sum_j M_j^S \otimes |j\rangle^{S'}. \tag{5.77}$$

Appying this isometry to a density operator ρ gives the following state

$$U^{S\to SS'}(\rho) = \sum_{j,j'} M_j^S \rho (M_{j'}^S)^\dagger \otimes |j\rangle\langle j'|^{S'}. \tag{5.78}$$

We can then apply a von Neumann measurement with projection operators $\{|j\rangle\langle j|\}_j$ to the system S', which then gives the following post-measurement state:

$$\frac{(I^S \otimes |j\rangle\langle j|^{S'})(U^{S\to SS'}(\rho))(I^S \otimes |j\rangle\langle j|^{S'})}{\text{Tr}\Big\{(I^S \otimes |j\rangle\langle j|^{S'})(U^{S\to SS'}(\rho))\Big\}} = \frac{M_j^S \rho (M_j^S)^\dagger}{\text{Tr}\big\{(M_j^S)^\dagger M_j^S \rho\big\}} \otimes |j\rangle\langle j|^{S'}. \tag{5.79}$$

The result is then the same as that in (5.75).

Exercise 5.4.1 Suppose that there is a set of density operators ρ_k^S and a POVM $\left\{\Lambda_k^S\right\}$ that identifies these states with high probability, in the sense that

$$\forall k \quad \text{Tr}\big\{\Lambda_k^S \rho_k^S\big\} \geq 1 - \epsilon, \tag{5.80}$$

where ϵ is some small number such that $\epsilon > 0$. Construct a coherent measurement $U^{S\to SS'}$ and show that the coherent measurement has a high probability of success in the sense that

$$\langle \phi_k|^{RS} \langle k|^{S'} U^{S\to SS'} |\phi_k\rangle^{RS} \geq 1 - \epsilon, \tag{5.81}$$

where each $|\phi_k\rangle^{RS}$ is a purification of ρ_k.

5.5 History and Further Reading

The purified view of quantum mechanics has long been part of quantum information theory (e.g., see the book of Nielsen and Chuang [197] or Yard's thesis [266]). The notion of an isometric extension of a quantum channel is due to early work of Stinespring [236]. Giovannetti and Fazio discussed some of the observations about the amplitude damping that appear in our exercises [102]. Devetak and Shor introduced generalized dephasing channels in the context of trade-off coding and they also introduced the notion of a degradable quantum channel [73]. King *et al.* studied the quantum Hadamard channels in Ref. [172]. Coherent instruments and measurements appeared in Refs. [68, 75, 156] as part of the decoder used in several quantum coding theorems. We exploit them in Chapters 23 and 24.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Part III

Unit Quantum Protocols

CHAPTER 6

Three Unit Quantum Protocols

This chapter begins our first exciting application of the postulates of the quantum theory to quantum communication. We study the fundamental, unit quantum communication protocols. These protocols involve a single sender, whom we name Alice, and a single receiver, whom we name Bob. The protocols are ideal and noiseless because we assume that Alice and Bob can exploit perfect classical communication, perfect quantum communication, and perfect entanglement. At the end of this chapter, we suggest how to incorporate imperfections into these protocols for later study.

Alice and Bob may wish to perform one of several quantum information processing tasks, such as the transmission of classical information, quantum information, or entanglement. Several fundamental protocols make use of these resources:

1. We will see that noiseless entanglement is an important resource in quantum Shannon theory because it enables Alice and Bob to perform other protocols that are not possible with classical resources only. We will present a simple, idealized protocol for generating entanglement, named *entanglement distribution*.

2. Alice may wish to communicate classical information to Bob. A trivial method, named *elementary coding*, is a simple way for doing so and we discuss it briefly.

3. A more elegant technique for transmitting classical information is *super-dense coding*. It exploits a noiseless qubit channel and shared entanglement to transmit more classical information than would be possible with a noiseless qubit channel only.

4. Finally, Alice may wish to transmit quantum information to Bob. A trivial method for Alice to transmit quantum information is for her to exploit a noiseless qubit channel. Though, it is useful to have other ways for transmitting quantum information because such a resource is difficult to engineer in practice. An alternative, surprising method for transmitting quantum information is *quantum teleportation*. The teleportation protocol exploits classical communication and shared entanglement to transmit quantum information.

Each of these protocols is a fundamental unit protocol and provides a foundation for asking further questions in quantum Shannon theory. In fact, the discovery of these latter two protocols was the stimulus for much of the original research in quantum Shannon theory.

We introduce the technique of *resource counting* in this chapter. This technique is of practical importance because it quantifies the communication cost of achieving a certain task. We include only nonlocal resources in a resource count—nonlocal resources include classical or quantum communication or shared entanglement.

It is important to minimize the use of certain resources, such as noiseless entanglement or a noiseless qubit channel, in a given protocol because they are expensive. Given a certain implementation of a quantum information processing task, we may wonder if there is a way of implementing it that consumes fewer resources. A proof that a given protocol is the best that we can hope to do is an optimality proof (also known as a converse proof, as discussed in Section 2.1.3). We argue, based on good physical grounds, that the protocols in this chapter are the best implementations of the desired quantum information processing task.

6.1 Nonlocal Unit Resources

We first briefly define what we mean by a noiseless qubit channel, a noiseless classical bit channel, and noiseless entanglement. Each of these resources is a *nonlocal, unit resource*. A resource is *nonlocal* if two spatially separated parties share it or if one generates it to the other. We say that a resource is *unit* if it comes in some "gold standard" form, such as qubits, classical bits, or entangled bits. It is important to establish these definitions so that we can check whether a given protocol is truly simulating one of these resources.

A noiseless qubit channel is any mechanism that implements the following map:

$$|i\rangle^A \rightarrow |i\rangle^B, \tag{6.1}$$

where $i \in \{0,1\}$, $\{|0\rangle^A, |1\rangle^A\}$ is some preferred orthonormal basis on Alice's system, and $\{|0\rangle^B, |1\rangle^B\}$ is some preferred orthonormal basis on Bob's system. The bases do not have to be the same, but it must be clear which basis each party is using. The above map is linear so that it preserves arbitrary superposition states (it preserves any qubit). For example, the map acts as follows on a superposition state:

$$\alpha|0\rangle^A + \beta|1\rangle^A \rightarrow \alpha|0\rangle^B + \beta|1\rangle^B. \tag{6.2}$$

We can also write it as the following isometry:

$$\sum_{i=0}^{1} |i\rangle^B\langle i|^A. \tag{6.3}$$

Any information processing protocol that implements the above map simulates a noiseless qubit channel. We label the communication resource of a noiseless qubit channel as follows:

$$[q \rightarrow q], \tag{6.4}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where the notation indicates one forward use of a noiseless qubit channel.

A noiseless classical bit channel is any mechanism that implements the following map:

$$|i\rangle\langle i|^A \rightarrow |i\rangle\langle i|^B, \tag{6.5}$$

$$|i\rangle\langle j|^A \rightarrow 0 \quad \text{for } i \neq j \tag{6.6}$$

where $i, j \in \{0, 1\}$ and the orthonormal bases are again arbitrary. This channel maintains the diagonal elements of a density operator in the basis $\left\{|0\rangle^A, |1\rangle^A\right\}$, but it eliminates the off-diagonal elements. We can write it as the following map:

$$\rho \rightarrow \sum_{i=0}^{1} |i\rangle^B\langle i|^A \rho |i\rangle^A\langle i|^B. \tag{6.7}$$

This resource is weaker than a noiseless qubit channel because it does not require Alice and Bob to maintain arbitrary superposition states—it merely transfers classical information. Alice can of course use the above channel to transmit classical information to Bob. She can prepare either of the classical states $|0\rangle\langle 0|$ or $|1\rangle\langle 1|$, send it through the classical channel, and Bob performs a computational basis measurement to determine the message Alice transmits. We denote the communication resource of a noiseless classical bit channel as follows:

$$[c \rightarrow c], \tag{6.8}$$

where the notation indicates one forward use of a noiseless classical bit channel.

We can study other ways of transmitting classical information. For example, suppose that Alice flips a fair coin that chooses the state $|0\rangle^A$ or $|1\rangle^A$ with equal probability. The resulting state is the following density operator:

$$\frac{1}{2}\Big(|0\rangle\langle 0|^A + |1\rangle\langle 1|^A\Big). \tag{6.9}$$

Suppose that she sends the above state through a noiseless classical channel. The resulting density operator for Bob is as follows:

$$\frac{1}{2}\Big(|0\rangle\langle 0|^B + |1\rangle\langle 1|^B\Big). \tag{6.10}$$

The above classical bit channel map does not necessarily preserve off-diagonal elements of a density operator. Suppose instead that Alice prepares a superposition state

$$\frac{|0\rangle^A + |1\rangle^A}{\sqrt{2}}. \tag{6.11}$$

The density operator corresponding to this state is

$$\frac{1}{2}\Big(|0\rangle\langle 0|^A + |0\rangle\langle 1|^A + |1\rangle\langle 0|^A + |1\rangle\langle 1|^A\Big). \tag{6.12}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Suppose Alice then transmits this state through the above classical channel. The classical channel eliminates all the off-diagonal elements of the density operator and the resulting state for Bob is as follows:

$$\frac{1}{2}\Big(|0\rangle\langle 0|^B + |1\rangle\langle 1|^B\Big). \tag{6.13}$$

Thus, it is impossible for a noiseless classical channel to simulate a noiseless qubit channel because it cannot maintain arbitrary superposition states. Though, it is possible for a noiseless qubit channel to simulate a noiseless classical bit channel and we denote this fact with the following *resource inequality*:

$$[q \to q] \geq [c \to c]. \tag{6.14}$$

Noiseless quantum communication is therefore a stronger resource than noiseless classical communication.

Exercise 6.1.1 Show that the noisy dephasing channel in (4.206) with $p = 1/2$ is equivalent to a noiseless classical bit channel.

The final resource that we consider is shared entanglement. The ebit is our "gold standard" resource for pure bipartite (two-party) entanglement, and we will make this point more clear operationally in Chapter 18. An ebit is the following state of two qubits:

$$\left|\Phi^+\right\rangle^{AB} \equiv \frac{|00\rangle^{AB} + |11\rangle^{AB}}{\sqrt{2}}, \tag{6.15}$$

where Alice possesses the first qubit and Bob possesses the second.

Below, we show how a noiseless qubit channel can generate a noiseless ebit through a simple protocol named *entanglement distribution*. Though, an ebit cannot simulate a noiseless qubit channel (for reasons which we explain later). Therefore, noiseless quantum communication is the strongest of all three resources, and entanglement and classical communication are in some sense "orthogonal" to one another because neither can simulate the other.

6.2 Protocols

6.2.1 Entanglement Distribution

The entanglement distribution protocol is the most basic of the three unit protocols. It exploits one use of a noiseless qubit channel to establish one shared noiseless ebit. It consists of the following two steps:

1. Alice prepares a Bell state locally in her laboratory. She prepares two qubits in the state $|0\rangle^A|0\rangle^{A'}$, where we label the first qubit as A and the second qubit as A'. She performs a Hadamard gate on qubit A to produce the following state:

$$\left(\frac{|0\rangle^A + |1\rangle^A}{\sqrt{2}}\right)|0\rangle^{A'}. \tag{6.16}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

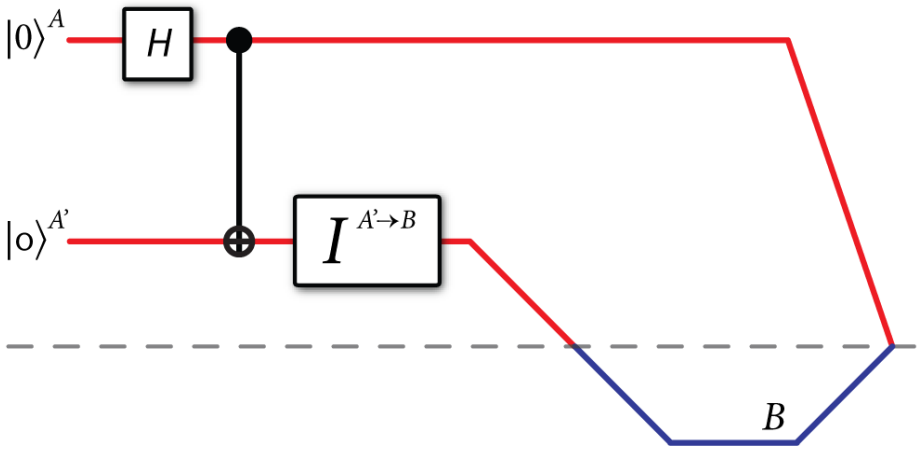

Figure 6.1: The above figure depicts a protocol for entanglement distribution. Alice performs local operations (the Hadamard and CNOT) and consumes one use of a noiseless qubit channel to generate one noiseless ebit $|\Phi^+\rangle^{AB}$ shared with Bob.

She then performs a CNOT gate with qubit A as the source qubit and qubit A' as the target qubit. The state becomes the following Bell state:

$$|\Phi^+\rangle^{AA'} = \frac{|00\rangle^{AA'} + |11\rangle^{AA'}}{\sqrt{2}}. \tag{6.17}$$

2. She sends qubit A' to Bob with one use of a noiseless qubit channel. Alice and Bob then share the ebit $|\Phi^+\rangle^{AB}$.

Figure 6.1 depicts the entanglement distribution protocol.

The following resource inequality quantifies the nonlocal resources consumed or generated in the above protocol:

$$[q \to q] \geq [qq], \tag{6.18}$$

where $[q \to q]$ denotes one forward use of a noiseless qubit channel and $[qq]$ denotes a shared, noiseless ebit. The meaning of the resource inequality is that there exists a protocol that consumes the resource on the left in order to generate the resource on the right. The best analogy is to think of a resource inequality as a "chemical reaction"-like formula, where the protocol is like a chemical reaction that transforms one resource into another.

There are several subtleties to notice about the above protocol and its corresponding resource inequality:

1. We are careful with the language when describing the resource state. We described the state $|\Phi^+\rangle$ as a Bell state in the first step because it is a local state in Alice's laboratory. We only used the term "ebit" to describe the state after the second step, when the state becomes a nonlocal resource shared between Alice and Bob.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

2. The resource count involves nonlocal resources only—we do not factor any local operations, such as the Hadamard gate or the CNOT gate, into the resource count. This line of thinking is different from the theory of computation, where it is of utmost importance to minimize the number of steps involved in a computation. In this book, we are developing a theory of quantum communication and we thus count nonlocal resources only.

3. We are assuming that it is possible to perform all local operations perfectly. This line of thinking is another departure from practical concerns that one might have in fault tolerant quantum computation, the study of the propagation of errors in quantum operations. Performing a CNOT gate is a highly nontrivial task at the current stage of experimental development in quantum computation, with most implementations being far from perfect. Nevertheless, we proceed forward with this communication-theoretic line of thinking.

The following exercises outline classical information processing tasks that are analogous to the task of entanglement distribution.

Exercise 6.2.1 Outline a protocol for *common randomness distribution*. Suppose that Alice and Bob have available one use of a noiseless classical bit channel. Give a method for them to implement the following resource inequality:

$$[c \to c] \geq [cc], \tag{6.19}$$

where $[c \to c]$ denotes one forward use of a noiseless classical bit channel and $[cc]$ denotes a shared, nonlocal bit of common randomness.

Exercise 6.2.2 Consider three parties Alice, Bob, and Eve and suppose that a noiseless private channel connects Alice to Bob. Privacy here implies that Eve does not learn anything about the information that traverses the private channel—Eve's probability distribution is independent of Alice and Bob's:

$$p_{A,B,E}(a,b,e) = p_A(a)p_{B|A}(b|a)p_E(e). \tag{6.20}$$

For a noiseless private bit channel, $p_{B|A}(b|a) = \delta_{b,a}$. A noiseless secret key corresponds to the following distribution:

$$p_{A,B,E}(a,b,e) = \frac{1}{2}\delta_{b,a}p_E(e), \tag{6.21}$$

where $\frac{1}{2}$ implies that the key is equal to '0' or '1' with equal probability, $\delta_{b,a}$ implies a perfectly correlated secret key, and the factoring of the distribution $p_{A,B,E}(a,b,e)$ implies the secrecy of the key (Eve's information is independent of Alice and Bob's). The difference between a noiseless private bit channel and a noiseless secret key is that the private channel is a dynamic resource while the secret key is a shared, static resource. Show that it is possible to upgrade the protocol for common randomness distribution to a protocol for *secret key*

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

distribution, if Alice and Bob share a noiseless private bit channel. That is, show that they can achieve the following resource inequality:

$$[c \to c]_{\text{priv}} \geq [cc]_{\text{priv}}, \tag{6.22}$$

where $[c \to c]_{\text{priv}}$ denotes one forward use of a noiseless private bit channel and $[cc]_{\text{priv}}$ denotes one bit of shared, noiseless secret key.

Entanglement and Quantum Communication

Can entanglement enable two parties to communicate quantum information? It is natural to wonder if there is a protocol corresponding to the following resource inequality:

$$[qq] \geq [q \to q]. \tag{6.23}$$

Such a resource inequality would be of great utility.

Unfortunately, it is physically impossible to construct a protocol that implements the above resource inequality. The argument against such a protocol arises from the theory of relativity. Specifically, the theory of relativity prohibits information transfer or signaling at a speed greater than the speed of light. Suppose that two parties share noiseless entanglement over a large distance. That resource is a static resource, possessing only shared quantum correlations. If a protocol were to exist that implements the above resource inequality, it would imply that two parties could communicate quantum information faster than the speed of light, because they would be exploiting the entanglement for the instantaneous transfer of quantum information.

The entanglement distribution resource inequality is only "one-way," as in (6.18). Quantum communication is therefore strictly stronger than shared entanglement when no other nonlocal resources are available.

6.2.2 Elementary Coding

We can also send classical information with a noiseless qubit channel. A simple protocol for doing so is *elementary coding.* This protocol consists of the following steps:

1. Alice prepares either $|0\rangle$ or $|1\rangle$, depending on the classical bit that she would like to send.

2. She transmits this state over the noiseless qubit channel and Bob receives the qubit.

3. Bob performs a measurement in the computational basis to determine the classical bit that Alice transmitted.

Elementary coding succeeds without error because Bob's measurement can always distinguish the classical states $|0\rangle$ and $|1\rangle$. The following resource inequality applies to elementary coding:

$$[q \to q] \geq [c \to c]. \tag{6.24}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Again, we are only counting nonlocal resources in the resource count—we do not count the state preparation at the beginning or the measurement at the end.

If no other resources are available for consumption, the above resource inequality is optimal—one cannot do better than to transmit one classical bit of information per use of a noiseless qubit channel. This result may be a bit frustrating at first, because it may seem that we could exploit the continuous degrees of freedom in the probability amplitudes of a qubit state for encoding more than one classical bit per qubit. Unfortunately, there is no way that we can access the information in the continuous degrees of freedom with any measurement scheme. The result of Exercise 4.2.2 demonstrates the optimality of the above protocol, and it holds as well by use of the Holevo bound in Chapter 11.

6.2.3 Quantum Super-Dense Coding

We now outline a protocol named *super-dense coding*. It is named such because it has the striking property that noiseless entanglement can double the classical communication ability of a noiseless qubit channel. It consists of three steps:

1. Suppose that Alice and Bob share an ebit $|\Phi^+\rangle^{AB}$. Alice applies one of four unitary operations $\{I, X, Z, XZ\}$ to her side of the above state. The state becomes one of the following four Bell states (up to a global phase), depending on the message that Alice chooses:
$$\left|\Phi^+\right\rangle^{AB}, \qquad \left|\Phi^-\right\rangle^{AB}, \qquad \left|\Psi^+\right\rangle^{AB}, \qquad \left|\Psi^-\right\rangle^{AB}. \tag{6.25}$$
The definitions of these Bell states are in (3.194-3.195).

2. She transmits her qubit to Bob with one use of a noiseless qubit channel.

3. Bob performs a Bell measurement (a measurement in the basis $\{|\Phi^+\rangle^{AB}, |\Phi^-\rangle^{AB}, |\Psi^+\rangle^{AB}, |\Psi^-\rangle^{AB}\}$) to distinguish perfectly the four states—he can distinguish the states because they are all orthogonal to each other.

Thus, Alice can transmit two classical bits (corresponding to the four messages) if she shares a noiseless ebit with Bob and uses a noiseless qubit channel. Figure 6.2 depicts the protocol for quantum super-dense coding.

The super-dense coding protocol implements the following resource inequality:

$$[qq] + [q \rightarrow q] \geq 2[c \rightarrow c]. \tag{6.26}$$

Notice again that the resource inequality counts the use of nonlocal resources only—we do not count the local operations at the beginning of the protocol or the Bell measurement at the end of the protocol.

Also, notice that we could have implemented two noiseless classical bit channels with two instances of elementary coding:

$$2[q \rightarrow q] \geq 2[c \rightarrow c]. \tag{6.27}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

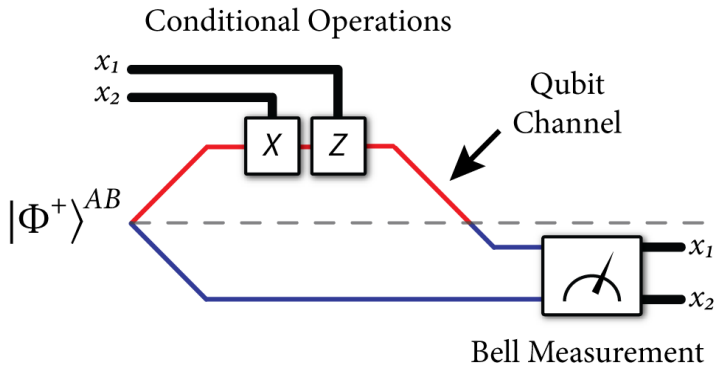

Figure 6.2: The above figure depicts the dense coding protocol. Alice and Bob share an ebit before the protocol begins. Alice would like to transmit two classical bits x_1x_2 to Bob. She performs a Pauli rotation conditional on her two classical bits and sends her half of the ebit over a noiseless qubit channel. Bob can then recover the two classical bits by performing a Bell measurement.

Though, this method is not as powerful as the super-dense coding protocol—in super-dense coding, we consume the weaker resource of an ebit to help transmit two classical bits, instead of consuming the stronger resource of an extra noiseless qubit channel.

The dense coding protocol also transmits the classical bits *privately*. Suppose a third party intercepts the qubit that Alice transmits. There is no measurement that the third party can perform to determine which message Alice transmits because the local density operator of all of the Bell states is the same and equal to the maximally mixed state π^A (the information for the eavesdropper is constant for each message that Alice transmits). The privacy of the protocol is due to Alice and Bob sharing maximal entanglement. We exploit this aspect of the dense coding protocol when we make it coherent in Chapter 7.

6.2.4 Quantum Teleportation

Perhaps the most striking protocol in noiseless quantum communication is the *quantum teleportation protocol*. The protocol destroys the quantum state of a qubit in one location and recreates it on a qubit at a distant location, with the help of shared entanglement. Thus, the name "teleportation" corresponds well to the mechanism that occurs.

The teleportation protocol is actually a flipped version of the super-dense coding protocol, in the sense that Alice and Bob merely "swap their equipment." The first step in understanding teleportation is to perform a few algebraic steps using the tricks of the tensor product and the Bell state substitutions from Exercise 3.5.13. Consider a qubit $|\psi\rangle^{A'}$ that Alice possesses, where

$$|\psi\rangle^{A'} \equiv \alpha|0\rangle^{A'} + \beta|1\rangle^{A'}. \tag{6.28}$$

Suppose she shares a maximally entangled state $|\Phi^+\rangle^{AB}$ with Bob. The joint state of the systems A', A, and B is as follows:

$$|\psi\rangle^{A'}|\Phi^+\rangle^{AB}. \tag{6.29}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

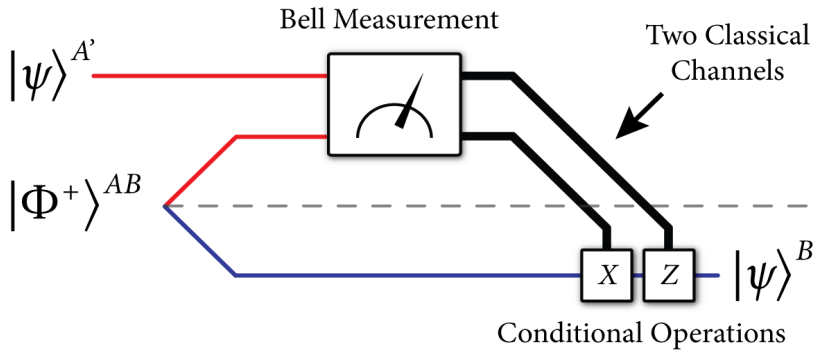

Figure 6.3: The above figure depicts the teleportation protocol. Alice would like to transmit an arbitrary quantum state $|\psi\rangle^{A'}$ to Bob. Alice and Bob share an ebit before the protocol begins. Alice can "teleport" her quantum state to Bob by consuming the entanglement and two uses of a noiseless classical bit channel.

Let us first explicitly write out this state:

$$|\psi\rangle^{A'}|\Phi^+\rangle^{AB} = \left(\alpha|0\rangle^{A'} + \beta|1\rangle^{A'}\right)\left(\frac{|00\rangle^{AB} + |11\rangle^{AB}}{\sqrt{2}}\right). \tag{6.30}$$

Distributing terms gives the following equality:

$$= \frac{1}{\sqrt{2}}\left[\alpha|000\rangle^{A'AB} + \beta|100\rangle^{A'AB} + \alpha|011\rangle^{A'AB} + \beta|111\rangle^{A'AB}\right]. \tag{6.31}$$

We use the relations in Exercise 3.5.13 to rewrite the joint system $A'A$ in the Bell basis:

$$= \frac{1}{2}\begin{bmatrix} \alpha\left(|\Phi^+\rangle^{A'A} + |\Phi^-\rangle^{A'A}\right)|0\rangle^B + \beta\left(|\Psi^+\rangle^{A'A} - |\Psi^-\rangle^{A'A}\right)|0\rangle^B \\ +\alpha\left(|\Psi^+\rangle^{A'A} + |\Psi^-\rangle^{A'A}\right)|1\rangle^B + \beta\left(|\Phi^+\rangle^{A'A} - |\Phi^-\rangle^{A'A}\right)|1\rangle^B \end{bmatrix} \tag{6.32}$$

Simplifying gives the following equivalence:

$$= \frac{1}{2}\begin{bmatrix} |\Phi^+\rangle^{A'A}\left(\alpha|0\rangle^B + \beta|1\rangle^B\right) + |\Phi^-\rangle^{A'A}\left(\alpha|0\rangle^B - \beta|1\rangle^B\right) \\ +|\Psi^+\rangle^{A'A}\left(\alpha|1\rangle^B + \beta|0\rangle^B\right) + |\Psi^-\rangle^{A'A}\left(\alpha|1\rangle^B - \beta|0\rangle^B\right) \end{bmatrix}. \tag{6.33}$$

We can finally rewrite the state as four superposition terms, with a distinct Pauli operator applied to Bob's system B for each term in the superposition:

$$= \frac{1}{2}\left[|\Phi^+\rangle^{A'A}|\psi\rangle^B + |\Phi^-\rangle^{A'A}Z|\psi\rangle^B + |\Psi^+\rangle^{A'A}X|\psi\rangle^B + |\Psi^-\rangle^{A'A}XZ|\psi\rangle^B\right]. \tag{6.34}$$

We now outline the three steps of the teleportation protocol (Figure 6.3 depicts the protocol):

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

1. Alice performs a Bell measurement on her systems $A'A$. The state collapses to one of the following four states with uniform probability:

$$\left|\Phi^{+}\right\rangle^{A'A}|\psi\rangle^{B}, \tag{6.35}$$
$$\left|\Phi^{-}\right\rangle^{A'A}Z|\psi\rangle^{B}, \tag{6.36}$$
$$\left|\Psi^{+}\right\rangle^{A'A}X|\psi\rangle^{B}, \tag{6.37}$$
$$\left|\Psi^{-}\right\rangle^{A'A}XZ|\psi\rangle^{B}. \tag{6.38}$$

   Notice that the state resulting from the measurement is a product state with respect to the cut $A'A-B$, regardless of the outcome of the measurement. At this point, Alice knows whether Bob's state is $|\psi\rangle^{B}$, $Z|\psi\rangle^{B}$, $X|\psi\rangle^{B}$, or $XZ|\psi\rangle^{B}$ because she knows the result of the measurement. On the other hand, Bob does not know anything about the state of his system B—Exercise 4.4.9 proves that his local density operator is the maximally mixed state π^{B} just after Alice performs the measurement. Thus, there is no teleportation of quantum information at this point because Bob's state is completely independent of the original state $|\psi\rangle$. In other words, teleportation cannot be instantaneous.

2. Alice transmits two classical bits to Bob that indicate which of the four measurement results she obtains. After Bob receives the classical information, he is immediately certain which operation he needs to perform in order to restore his state to Alice's original state $|\psi\rangle$. Notice that he does not need to have knowledge of the state in order to restore it—he only needs knowledge of the restoration operation.

3. Bob performs the restoration operation: one of the identity, a Pauli X operator, a Pauli Z operator, or the Pauli operator XZ, depending on the classical information that he receives from Alice.

Teleportation is an *oblivious* protocol because Alice and Bob do not require any knowledge of the quantum state being teleported in order to perform it. We might also say that this feature of teleportation makes it universal—it works independent of the input state.

You might think that the teleportation protocol violates the no-cloning theorem because a "copy" of the state appears on Bob's system. But this violation does not occur at any point in the protocol because the Bell measurement destroys the information about the state of Alice's original information qubit while recreating it somewhere else. Also, notice that the result of the Bell measurement is independent of the particular probability amplitudes α and β corresponding to the state Alice wishes to teleport.

The teleportation protocol is not an instantaneous teleportation, as portrayed in the television episodes of Star Trek. There is no transfer of quantum information instantaneously after the Bell measurement because Bob's local description of the B system is the maximally mixed state π. It is only after he receives the classical bits to "telecorrect" his state that the transfer occurs. It must be this way—otherwise, they would be able to communicate

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

faster than the speed of light, and superluminal communication is not allowed in the theory of relativity.

Finally, we can phrase the teleportation protocol as a resource inequality:

$$[qq] + 2[c \to c] \geq [q \to q]. \tag{6.39}$$

Again, we factor only nonlocal resources into the resource count. The above resource inequality is perhaps the most surprising of the three unit protocols we have studied so far. It combines two resources, noiseless entanglement and noiseless classical communication, that achieve noiseless quantum communication even though they are both individually weaker than it. This protocol and super-dense coding are two of the most fundamental protocols in quantum communication theory because they sparked the notion that there are clever ways of combining resources to generate other resources.

In Exercise 6.2.3 below, we discuss a variation of teleportation called *remote state preparation*, where Alice possesses a classical description of the state that she wishes to teleport. With this knowledge, it is possible to reduce the amount of classical communication necessary for teleportation.

Exercise 6.2.3 *Remote state preparation* is a variation on the teleportation protocol. We consider a simple example of a remote state preparation protocol. Suppose Alice possesses a classical description of a state $|\psi\rangle \equiv \left(|0\rangle + e^{i\phi}|1\rangle\right)/\sqrt{2}$ (on the equator of the Bloch sphere) and she shares an ebit $|\Phi^+\rangle^{AB}$ with Bob. Alice would like to prepare this state on Bob's system. Show that Alice can prepare this state on Bob's system if she measures her system A in the $\left\{|\psi^*\rangle, |\psi^{\perp *}\rangle\right\}$ basis, transmits one classical bit, and Bob performs a recovery operation conditional on the classical information. (Note that $|\psi^*\rangle$ is the conjugate of the vector $|\psi\rangle$).

Exercise 6.2.4 *Third-party controlled teleportation* is another variation on the teleportation protocol. Suppose that Alice, Bob, and Charlie possess a GHZ state:

$$|\Phi_{\text{GHZ}}\rangle \equiv \frac{|000\rangle^{ABC} + |111\rangle^{ABC}}{\sqrt{2}}. \tag{6.40}$$

Alice would like to teleport an arbitrary qubit to Bob. She performs the usual steps in the teleportation protocol. Give the final steps that Charlie should perform and the information that he should transmit to Bob in order to complete the teleportation protocol. (Hint: The resource inequality for the protocol is as follows:

$$[qqq]_{ABC} + 2[c \to c]_{A\to B} + [c \to c]_{C\to B} \geq [q \to q]_{A\to B}, \tag{6.41}$$

where $[qqq]_{ABC}$ represents the resource of the GHZ state shared between Alice, Bob, and Charlie, and the other resources are as before with the directionality of communication indicated by the corresponding subscript.)

Exercise 6.2.5 *Gate teleportation* is yet another variation of quantum teleportation that is useful in fault-tolerant quantum computation. Suppose that Alice would like to perform a

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

single-qubit gate U on a qubit in state $|\psi\rangle$. Suppose that the gate U is difficult to perform, but that $U\sigma_i U^\dagger$, where σ_i is one of the single-qubit Pauli operators, is much less difficult to perform. A protocol for gate teleportation is as follows. Alice and Bob first prepare the ebit $U^B|\Phi^+\rangle^{AB}$. Alice performs a Bell measurement on her qubit $|\psi\rangle^{A'}$ and system A. She transmits two classical bits to Bob and Bob performs one of the four corrective operations $U\sigma_i U^\dagger$ on his qubit. Show that this protocol works, i.e., Bob's final state is $U|\psi\rangle$.

Exercise 6.2.6 Show that it is possible to simulate a dephasing qubit channel by the following technique. First, Alice prepares a maximally entangled Bell state $|\Phi^+\rangle$. She sends half of it to Bob through a dephasing qubit channel. She and Bob perform the usual teleportation protocol. Show that this procedure gives the same result as sending a qubit through a dephasing channel. (Hint: This result holds because the dephasing channel commutes with all Pauli operators.)

Exercise 6.2.7 Construct an *entanglement swapping protocol* from the teleportation protocol. That is, suppose that Charlie and Alice possess a bipartite state $|\psi\rangle^{CA}$. Show that if Alice teleports her half of the state $|\psi\rangle^{CA}$ to Bob, then Charlie and Bob share the state $|\psi\rangle^{CB}$. A special case of this protocol is when the state $|\psi\rangle^{CA}$ is an ebit. Then the protocol is equivalent to an entanglement swapping protocol.

6.3 Optimality of the Three Unit Protocols

We now consider several arguments that may seem somewhat trivial at first, but they are crucial in a good theory of quantum communication. We are always thinking about the optimality of certain protocols—if there is a better, cheaper way to perform a given protocol, we would prefer to do it this way so that we do not have to pay expensive bills to the quantum communication companies of the future.[1] There are several questions that we can ask about the above protocols:

1. In entanglement distribution, is one ebit per qubit the best that we can do, or is it possible to generate more than one ebit with a single use of a noiseless qubit channel?

2. In super-dense coding, is it possible to generate two noiseless classical bit channels with less than one noiseless qubit channel or less than one noiseless ebit? Is it possible to generate more than two classical bit channels with the given resources?

3. In teleportation, is it possible to teleport more than one qubit with the given resources? Is it possible to teleport using less than two classical bits or less than one ebit?

In this section, we answer all these questions in the negative—all the protocols as given are optimal protocols. Here, we begin to see the beauty of the resource inequality formalism.

[1] If you will be working at a future quantum communication company, it also makes sense to find optimal protocols so that you can squeeze in more customers with your existing physical resources!

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

It allows us to chain protocols together to make new protocols. We exploit this idea in the forthcoming optimality arguments.

First, let us tackle the optimality of entanglement distribution. Is there a protocol that implements any other resource inequality such as

$$[q \to q] \geq E[qq], \tag{6.42}$$

where the rate E of entanglement generation is greater than one?

We show that such a resource inequality can never occur, i.e., it is optimal for $E = 1$. Suppose such a resource inequality with $E > 1$ does exist. Under an assumption of free forward classical communication, we can combine the above resource inequality with teleportation to achieve the following resource inequality:

$$[q \to q] \geq E[q \to q]. \tag{6.43}$$

We could then simply keep repeating this protocol to achieve an unbounded amount of quantum communication, which is impossible. Thus, it must be that $E = 1$.

Next, we consider the optimality of super-dense coding. We again exploit a proof by contradiction argument. Let us suppose that we have an unlimited amount of entanglement available. Suppose that there exists some "super-duper" dense coding protocol that generates an amount of classical communication greater than super-dense coding generates. That is, the classical communication output of super-duper-dense coding is $2C$ where $C > 1$, and its resource inequality is

$$[q \to q] + [qq] \geq 2C[c \to c]. \tag{6.44}$$

Then this super-duper-dense coding scheme (along with the infinite entanglement) gives the following resource inequality:

$$[q \to q] + \infty[qq] \geq 2C[c \to c] + \infty[qq]. \tag{6.45}$$

An infinite amount of entanglement is still available after executing the super-duper-dense coding protocol because it consumes only a finite amount of entanglement. We can then chain the above protocol with teleportation and achieve the following resource inequality:

$$2C[c \to c] + \infty[qq] \geq C[q \to q] + \infty[qq]. \tag{6.46}$$

Overall, we have then shown a scheme that achieves the following resource inequality:

$$[q \to q] + \infty[qq] \geq C[q \to q] + \infty[qq]. \tag{6.47}$$

We can continue with this protocol and perform it k times so that we implement the following resource inequality:

$$[q \to q] + \infty[qq] \geq C^k[q \to q] + \infty[qq]. \tag{6.48}$$

The result of this construction is that one noiseless qubit channel and an infinite amount of entanglement can generate an infinite amount of quantum communication. This result

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

is impossible physically because entanglement does not boost the capacity of a noiseless qubit channel. Also, the scheme is exploiting just one noiseless qubit channel along with the entanglement to generate an unbounded amount of quantum communication—it must be signaling superluminally in order to do so. Thus, the rate of classical communication in super-dense coding is optimal.

We leave the optimality arguments for teleportation as an exercise because they are similar to those for the super-dense coding protocol. Note that it is possible to prove optimality of these protocols without assumptions such as free classical communication (for the case of entanglement distribution), and we do so in Chapter 8.

Exercise 6.3.1 Show that it is impossible for $C > 1$ in the teleportation protocol where C is with respect to the following resource inequality:

$$2[c \to c] + [qq] \geq C[q \to q]. \tag{6.49}$$

Exercise 6.3.2 Show that the rates of the consumed resources in the teleportation and super-dense coding protocols are optimal.

6.4 Extensions for Quantum Shannon Theory

The previous section sparked some good questions that we might ask as a quantum Shannon theorist. We might also wonder what types of communication rates are possible if some of the consumed resources are noisy, rather than being perfect resources. We list some of these questions below.

Let us first consider entanglement distribution. Suppose that the consumed noiseless qubit channel in entanglement distribution is instead a noisy quantum channel $\mathcal{N}$ where $\mathcal{N}$ is some CPTP map. The communication task is then known as *entanglement generation*. We can rephrase the communication task as the following resource inequality:

$$\langle \mathcal{N} \rangle \geq E[qq]. \tag{6.50}$$

The meaning of the resource inequality is that we consume the resource of a noisy quantum channel $\mathcal{N}$ in order to generate entanglement between a sender and receiver at some rate E. We will make the definition of a quantum Shannon-theoretic resource inequality more precise when we begin our formal study of quantum Shannon theory, but the above definition should be sufficient for now. The optimal rate of entanglement generation with the noisy quantum channel $\mathcal{N}$ is known as the entanglement generation capacity of $\mathcal{N}$. This task is intimately related to the quantum communication capacity of $\mathcal{N}$, and we discuss the connection further in Chapter 23.

Let us now turn to super-dense coding. Suppose that the consumed noiseless qubit channel in super-dense coding is instead a noisy quantum channel $\mathcal{N}$. The name for this task is then *entanglement-assisted classical communication*. The following resource inequality captures the corresponding communication task:

$$\langle \mathcal{N} \rangle + E[qq] \geq C[c \to c]. \tag{6.51}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The meaning of the resource inequality is that we consume a noisy quantum channel $\mathcal{N}$ and noiseless entanglement at some rate E to produce noiseless classical communication at some rate C. We will study this protocol in depth in Chapter 20. We can also consider the scenario where the entanglement is no longer noiseless, but it is rather a general bipartite state ρ^{AB} that Alice and Bob share. The task is then known as noisy super-dense coding.[2] We study noisy super-dense coding in Chapter 21. The corresponding resource inequality is as follows (its meaning should be clear at this point):

$$\langle \rho^{AB} \rangle + Q[q \to q] \geq C[c \to c]. \tag{6.52}$$

We can ask the same questions for the teleportation protocol as well. Suppose that the entanglement resource is instead a noisy bipartite state ρ^{AB}. The task is then *noisy teleportation* and has the following resource inequality:

$$\langle \rho^{AB} \rangle + C[c \to c] \geq Q[q \to q]. \tag{6.53}$$

The questions presented in this section are some of the fundamental questions in quantum Shannon theory. We arrived at these questions simply by replacing the noiseless resources in the three fundamental noiseless protocols with noisy ones. We will spend a significant amount of effort building up our knowledge of quantum Shannon-theoretic tools that will be indispensable for answering these questions.

6.5 Three Unit Qudit Protocols

We end this chapter by studying the qudit versions of the three unit protocols. It is useful to have these versions of the protocols because we may want to process qudit systems with them.

The qudit resources are straightforward extensions of the qubit resources. A noiseless qudit channel is the following map:

$$|i\rangle^A \to |i\rangle^B, \tag{6.54}$$

where $\{|i\rangle^A\}_{i\in\{0,\ldots,d-1\}}$ is some preferred orthonormal basis on Alice's system and $\{|i\rangle^B\}_{i\in\{0,\ldots,d-1\}}$ is some preferred basis on Bob's system. We can also write the qudit channel map as the following isometry:

$$I^{A\to B} \equiv \sum_{i=0}^{d-1} |i\rangle^B \langle i|^A. \tag{6.55}$$

The map $I^{A\to B}$ preserves superposition states so that

$$\sum_{i=0}^{d-1} \alpha_i |i\rangle^A \to \sum_{i=0}^{d-1} \alpha_i |i\rangle^B. \tag{6.56}$$

[2]The name noisy super-dense coding could just as well apply to the former task of entanglement-assisted classical communication, but this terminology has "stuck" in the research literature for this specific quantum information processing task.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

A noiseless classical dit channel or *cdit* is the following map:

$$|i\rangle\langle i|^A \rightarrow |i\rangle\langle i|^B, \tag{6.57}$$

$$|i\rangle\langle j|^A \rightarrow 0 \text{ for } i \neq j. \tag{6.58}$$

A noiseless maximally entangled qudit state or an *edit* is as follows:

$$|\Phi\rangle^{AB} \equiv \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |i\rangle^A |i\rangle^B. \tag{6.59}$$

We quantify the "dit" resources with bit measures. For example, a noiseless qudit channel is the following resource:

$$\log d[q \rightarrow q], \tag{6.60}$$

where the logarithm is base two. Thus, one qudit channel can transmit $\log d$ qubits of quantum information so that the qubit remains our standard unit of quantum information. We quantify the amount of information transmitted according to the dimension of the space that is transmitted. For example, suppose that a quantum system has eight levels. We can then encode three qubits of quantum information in this eight-level system.

Likewise, a classical dit channel is the following resource:

$$\log d[c \rightarrow c], \tag{6.61}$$

so that a classical dit channel transmits $\log d$ classical bits. The parameter d here is the number of classical messages that the channel transmits.

Finally, an edit is the following resource:

$$\log d[qq]. \tag{6.62}$$

We quantify the amount of entanglement in a maximally entangled state by its Schmidt rank (see Theorem 3.6.1). We measure entanglement in units of ebits (we return to this issue in Chapter 18).

6.5.1 Entanglement Distribution

The extension of the entanglement distribution protocol to the qudit case is straightforward. Alice merely prepares the state $|\Phi\rangle^{AA'}$ in her laboratory and transmits the system A' through a noiseless qudit channel. She can prepare the state $|\Phi\rangle^{AA'}$ with two gates: the qudit analog of the Hadamard gate and the CNOT gate. The qudit analog of the Hadamard gate is the Fourier gate F introduced in Exercise 3.6.8 where

$$F : |l\rangle \rightarrow \frac{1}{\sqrt{d}} \sum_{j=0}^{d-1} \exp\left\{ \frac{2\pi i l j}{d} \right\} |j\rangle, \tag{6.63}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

so that

$$F \equiv \frac{1}{\sqrt{d}} \sum_{l,j=0}^{d-1} \exp\left\{ \frac{2\pi i l j}{d} \right\} |j\rangle\langle l|. \tag{6.64}$$

The qudit analog of the CNOT gate is the following controlled-shift gate:

$$\text{CNOT}_d \equiv \sum_{j=0}^{d-1} |j\rangle\langle j| \otimes X(j), \tag{6.65}$$

where $X(j)$ is defined in (3.209).

Exercise 6.5.1 Verify that Alice can prepare the maximally entangled qudit state $|\Phi\rangle^{AA'}$ locally by preparing $|0\rangle^A|0\rangle^{A'}$, applying F^A and CNOT_d. Show that

$$|\Phi\rangle^{AA'} \equiv \text{CNOT}_d \cdot F^A |0\rangle^A |0\rangle^{A'}. \tag{6.66}$$

The resource inequality for this qudit entanglement distribution protocol is as follows:

$$\log d[q \to q] \geq \log d[qq]. \tag{6.67}$$

6.5.2 Quantum Super-Dense Coding

The qudit version of the super-dense coding protocol proceeds analogously to the qubit case, with some notable exceptions. It still consists of three steps:

1. Alice and Bob begin with a maximally-entangled qudit state of the form:

$$|\Phi\rangle^{AB} = \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |i\rangle^A |i\rangle^B. \tag{6.68}$$

   Alice applies one of d^2 unitary operations in the set $\{X(x)Z(z)\}_{x,z=0}^{d-1}$ to her qudit. The shared state then becomes one of the d^2 maximally entangled qubit states in (3.240).

2. She sends her qudit to Bob with one use of a noiseless qudit channel.

3. Bob performs a measurement in the qudit Bell basis to determine the message Alice sent. The result of Exercise 3.6.11 is that these states are perfectly distinguishable with a measurement.

This qudit super-dense coding protocol implements the following resource inequality:

$$\log d[qq] + \log d[q \to q] \geq 2 \log d[c \to c]. \tag{6.69}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

6.5.3 Quantum Teleportation

The operations in the qudit teleportation protocol are again similar to the qubit case. The protocol proceeds in three steps:

1. Alice possesses an arbitrary qudit $|\psi\rangle^{A'}$ where

$$|\psi\rangle^{A'} \equiv \sum_{i=0}^{d-1} \alpha_i |i\rangle^{A'}. \tag{6.70}$$

   Alice and Bob share a maximally-entangled qudit state $|\Phi^+\rangle^{AB}$ of the form

$$|\Phi\rangle^{AB} = \frac{1}{\sqrt{d}} \sum_{j=0}^{d-1} |j\rangle^A |j\rangle^B. \tag{6.71}$$

   The joint state of Alice and Bob is then $|\psi\rangle^{A'}|\Phi\rangle^{AB}$. Alice performs a measurement in the basis $\{|\Phi_{i,j}\rangle^{A'A}\}_{i,j}$.

2. She transmits the measurement result i, j to Bob with the use of two classical dit channels.

3. Bob then applies the unitary transformation $Z^B(j)X^B(i)$ to his state to "telecorrect" it to Alice's original qudit.

We prove that this protocol works by analyzing the probability of the measurement result and the post-measurement state on Bob's system. The techniques that we employ here are different from those for the qubit case.

First, let us suppose that Alice would like to teleport the A' system of a state $|\psi\rangle^{RA'}$ that she shares with an inaccessible reference system R. This way, our teleportation protocol encompasses the most general setting in which Alice would like to teleport a mixed state on A'. Also, Alice shares the maximally entangled edit state $|\Phi\rangle^{AB}$ with Bob. Alice first performs a measurement of the systems A' and A in the basis $\{|\Phi_{i,j}\rangle^{A'A}\}_{i,j}$ where

$$|\Phi_{i,j}\rangle^{A'A} = U_{ij}^{A'} |\Phi\rangle^{A'A}, \tag{6.72}$$

and

$$U_{ij}^{A'} \equiv Z^{A'}(j)X^{A'}(i). \tag{6.73}$$

The measurement operators are thus

$$|\Phi_{i,j}\rangle\langle\Phi_{i,j}|^{A'A}. \tag{6.74}$$

Then the unnormalized post-measurement state is

$$|\Phi_{i,j}\rangle\langle\Phi_{i,j}|^{A'A} \; |\psi\rangle^{RA'}|\Phi\rangle^{AB}. \tag{6.75}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can rewrite this state as follows, by exploiting the definition of $|\Phi_{i,j}\rangle^{A'A}$ in (6.72):

$$|\Phi_{i,j}\rangle\langle\Phi|^{A'A} \left(U_{ij}^{\dagger}\right)^{A'} |\psi\rangle^{RA'}|\Phi\rangle^{AB}. \tag{6.76}$$

Recall the "Bell-state matrix identity" in Exercise 3.6.12 that holds for any maximally entangled state $|\Phi\rangle$. We can exploit this result to show that the action of the unitary $U_{ij}^{\dagger}$ on the A' system is the same as the action of the unitary U_{ij}^{*} on the A system:

$$|\Phi_{i,j}\rangle\langle\Phi|^{A'A} \left(U_{ij}^{*}\right)^{A} |\psi\rangle^{RA'}|\Phi\rangle^{AB}. \tag{6.77}$$

Then the unitary $\left(U_{ij}^{*}\right)^{A}$ commutes with the systems R and A':

$$|\Phi_{i,j}\rangle\langle\Phi|^{A'A} |\psi\rangle^{RA'} \left(U_{ij}^{*}\right)^{A}|\Phi\rangle^{AB}. \tag{6.78}$$

We can again apply the Bell state matrix identity in Exercise 3.6.12 to show that the state is equal to

$$|\Phi_{i,j}\rangle\langle\Phi|^{A'A} |\psi\rangle^{RA'} \left(U_{ij}^{\dagger}\right)^{B}|\Phi\rangle^{AB}. \tag{6.79}$$

Then we can commute the unitary $\left(U_{ij}^{\dagger}\right)^{B}$ all the way to the left, and we can switch the order of $|\psi\rangle^{RA'}$ and $|\Phi\rangle^{AB}$ without any problem because the system labels are sufficient to track the states in these systems:

$$\left(U_{ij}^{\dagger}\right)^{B}|\Phi_{i,j}\rangle\langle\Phi|^{A'A} |\Phi\rangle^{AB} |\psi\rangle^{RA'}. \tag{6.80}$$

Now let us consider the very special overlap $\langle\Phi|^{A'A} |\Phi\rangle^{AB}$ of the maximally entangled edit state with itself on different systems:

$$\begin{aligned}
\langle\Phi|^{A'A} |\Phi\rangle^{AB} &= \left(\frac{1}{\sqrt{d}}\sum_{i=0}^{d-1}\langle i|^{A'}\langle i|^{A}\right)\left(\frac{1}{\sqrt{d}}\sum_{j=0}^{d-1}|j\rangle^{A}|j\rangle^{B}\right) && (6.81)\\
&= \frac{1}{d}\sum_{i,j=0}^{d-1}\langle i|^{A'}\langle i|^{A}|j\rangle^{A}|j\rangle^{B} && (6.82)\\
&= \frac{1}{d}\sum_{i,j=0}^{d-1}\langle i|^{A'}\langle i|j\rangle^{A}|j\rangle^{B} && (6.83)\\
&= \frac{1}{d}\sum_{i=0}^{d-1}\langle i|^{A'}|i\rangle^{B} && (6.84)\\
&= \frac{1}{d}\sum_{i=0}^{d-1}|i\rangle^{B}\langle i|^{A'} && (6.85)\\
&= \frac{1}{d}I^{A'\to B}. && (6.86)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first equality follows by definition. The second equality follows from linearity and rearranging terms in the multiplication and summation. The third and fourth equalities follow by realizing that $\langle i|^A|j\rangle^A$ is an inner product and evaluating it for the orthonormal basis $\left\{|i\rangle^A\right\}$. The fifth equality follows by rearranging the bra and the ket. The final equality is our last important realization: the operator $\sum_{i=0}^{d-1}|i\rangle^B\langle i|^{A'}$ is the noiseless qudit channel $I^{A'\to B}$ that the teleportation protocol creates from the system A' to B (see the definition of a noiseless qudit channel in (6.55)). We might refer to this as the "teleportation isometry."

We now apply the teleportation isometry to the state in (6.80):

$$\left(U_{ij}^\dagger\right)^B|\Phi_{i,j}\rangle\langle\Phi|^{A'A}\ |\Phi\rangle^{AB}\ |\psi\rangle^{RA'} = \left(U_{ij}^\dagger\right)^B|\Phi_{i,j}\rangle^{A'A}\frac{1}{d}I^{A'\to B}\ |\psi\rangle^{RA'} \tag{6.87}$$

$$= \frac{1}{d}\left(U_{ij}^\dagger\right)^B|\Phi_{i,j}\rangle^{A'A}|\psi\rangle^{RB} \tag{6.88}$$

$$= \frac{1}{d}|\Phi_{i,j}\rangle^{A'A}\ \left(U_{ij}^\dagger\right)^B|\psi\rangle^{RB}. \tag{6.89}$$

We can compute the probability of receiving outcome i and j from the measurement when the input state is $|\psi\rangle^{RA'}$. It is just equal to the overlap of the above vector with itself:

$$p(i,j|\psi) = \left[\frac{1}{d}\langle\Phi_{i,j}|^{A'A}\ \langle\psi|^{RB}(U_{ij})^B\right]\left[\frac{1}{d}|\Phi_{i,j}\rangle^{A'A}\ \left(U_{ij}^\dagger\right)^B|\psi\rangle^{RB}\right] \tag{6.90}$$

$$= \frac{1}{d^2}\langle\Phi_{i,j}|^{A'A}|\Phi_{i,j}\rangle^{A'A}\ \langle\psi|^{RB}(U_{ij})^B\left(U_{ij}^\dagger\right)^B|\psi\rangle^{RB} \tag{6.91}$$

$$= \frac{1}{d^2}\langle\Phi_{i,j}|^{A'A}|\Phi_{i,j}\rangle^{A'A}\ \langle\psi|^{RB}|\psi\rangle^{RB} \tag{6.92}$$

$$= \frac{1}{d^2}. \tag{6.93}$$

Thus, the probability of the outcome (i,j) is completely random and independent of the input state. We would expect this to be the case for a universal teleportation protocol that operates independently of the input state. Thus, after normalization, the state on Alice and Bob's system is

$$|\Phi_{i,j}\rangle^{A'A}\ \left(U_{ij}^\dagger\right)^B|\psi\rangle^{RB}. \tag{6.94}$$

At this point, Bob does not know the result of the measurement. We obtain his density operator by tracing over the systems A', A, and R to which he does not have access and taking the expectation over all the measurement outcomes:

$$\mathrm{Tr}_{A'AR}\left\{\frac{1}{d^2}\sum_{i,j=0}^{d-1}|\Phi_{i,j}\rangle\langle\Phi_{i,j}|^{A'A}\ \left(U_{ij}^\dagger\right)^B|\psi\rangle\langle\psi|^{RB}(U_{ij})^B\right\}$$

$$= \frac{1}{d^2}\sum_{i,j=0}^{d-1}\left(U_{ij}^\dagger\right)^B\psi^B(U_{ij})^B \tag{6.95}$$

$$= \pi^B. \tag{6.96}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first equality follows by evaluating the partial trace and by defining

$$\psi^B \equiv \text{Tr}_R\Big\{|\psi\rangle\langle\psi|^{RB}\Big\}. \tag{6.97}$$

The second equality follows because applying a Heisenberg-Weyl operator uniformly at random completely randomizes a quantum state to be the maximally mixed state (see Exercise 4.4.9).

Now suppose that Alice sends the measurement results i and j over two uses of a noiseless classical dit channel. Bob then knows that the state is

$$\Big(U_{ij}^\dagger\Big)^B|\psi\rangle^{RB}, \tag{6.98}$$

and he can apply $(U_{ij})^B$ to make the overall state become $|\psi\rangle^{RB}$. This final step completes the teleportation process. The resource inequality for the qudit teleportation protocol is as follows:

$$\log d[qq] + 2\log d[c \to c] \geq \log d[q \to q]. \tag{6.99}$$

Exercise 6.5.2 Show that

$$\langle\Phi^+|^{A'A}\Big((U_{ij}^\dagger)^{A'}|\psi\rangle\langle\psi|^{A'}U_{ij}^{A'} \otimes \pi^A\Big)|\Phi^+\rangle^{A'A} = \frac{1}{d^2}\text{Tr}\Big\{(U_{ij}^\dagger)^{A'}|\psi\rangle\langle\psi|^{A'}U_{ij}^{A'}\Big\}. \tag{6.100}$$

Exercise 6.5.3 Show that

$$\Big((U_{ij}^\dagger)^B|\psi\rangle\langle\psi|^BU_{ij}^B\Big)U_{ij}^{A'}|\Phi^+\rangle\langle\Phi^+|^{A'A}\big|\Phi^+\big\rangle\big\langle\Phi^+\big|^{AB}|\Phi^+\rangle\langle\Phi^+|^{A'A}(U_{ij}^\dagger)^{A'}$$
$$= \frac{1}{d^2}\Big((U_{ij}^\dagger)^B|\psi\rangle\langle\psi|^BU_{ij}^B\Big) \otimes U_{ij}^{A'}|\Phi^+\rangle\langle\Phi^+|^{A'A}(U_{ij}^\dagger)^{A'}. \tag{6.101}$$

6.6 History and Further Reading

This chapter presented the three important protocols that exploit the three unit resources of classical communication, quantum communication, and entanglement. We learned, perhaps surprisingly, that it is possible to combine two resources together in interesting ways to simulate a different resource (in both super-dense coding and teleportation). These combinations of resources turn up quite a bit in quantum Shannon theory, and we see them in their most basic form in this chapter.

Bennett and Wiesner published the super-dense coding protocol in 1992 [35], and within a year, Bennett *et al.* realized that Alice and Bob could teleport particles if they swap their operations with respect to the super-dense coding protocol [23]. These two protocols were the seeds of much later work in quantum Shannon theory.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 7

Coherent Protocols

We introduced three protocols in the previous chapter: entanglement distribution, teleportation, and super-dense coding. The last two of these protocols, teleportation and super-dense coding, are perhaps more interesting than entanglement distribution because they demonstrate insightful ways that we can combine all three unit resources to achieve an information processing task.

It appears that teleportation and super-dense coding might be "inverse" protocols with respect to each other because teleportation arises from super-dense coding when Alice and Bob "swap" their equipment. But there is a fundamental asymmetry between these protocols when we consider their respective resource inequalities. Recall that the resource inequality for teleportation is

$$2[c \to c] + [qq] \geq [q \to q], \tag{7.1}$$

while that for super-dense coding is

$$[q \to q] + [qq] \geq 2[c \to c]. \tag{7.2}$$

The asymmetry in these protocols is that they are not *dual under resource reversal*. Two protocols are dual under resource reversal if the resources that one consumes are the same that the other generates and vice versa. Consider that the super-dense coding resource inequality in (7.2) generates two classical bit channels. Glancing at the left hand side of the teleportation resource inequality in (7.1), we see that two classical bit channels generated from super-dense coding are not sufficient to generate the noiseless qubit channel on the right hand side of (7.1)—the protocol requires the consumption of noiseless entanglement in addition to the consumption of the two noiseless classical bit channels.

Is there a way for teleportation and super-dense coding to become dual under resource reversal? One way is if we assume that *entanglement is a free resource*. This assumption is strong and we may have difficulty justifying it from a practical standpoint because noiseless entanglement is extremely fragile. It is also a powerful resource, as the teleportation and super-dense coding protocols demonstrate. But in the theory of quantum communication, we often make assumptions such as this one—such assumptions tend to give a dramatic simplification of a problem. Continuing with our development, let us assume that entanglement

is a free resource and that we do not have to factor it into the resource count. Under this assumption, the resource inequality for teleportation becomes

$$2[c \to c] \geq [q \to q], \tag{7.3}$$

and that for super-dense coding becomes

$$[q \to q] \geq 2[c \to c]. \tag{7.4}$$

Teleportation and super-dense coding are then dual under resource reversal under the "free-entanglement" assumption, and we obtain the following *resource equality*:

$$[q \to q] = 2[c \to c]. \tag{7.5}$$

Exercise 7.0.1 Suppose that the quantum capacity of a quantum channel assisted by an unlimited amount of entanglement is equal to some number Q. What is the capacity of that entanglement-assisted channel for transmitting classical information?

Exercise 7.0.2 How can we obtain the following resource equality? (Hint: Assume that some resource is free.)

$$[q \to q] = [qq]. \tag{7.6}$$

Which noiseless protocols did you use to show the above resource equality? The above resource equality is a powerful statement: entanglement and quantum communication are equivalent under the assumption that you have found.

Exercise 7.0.3 Suppose that the entanglement generation capacity of a quantum channel is equal to some number E. What is the quantum capacity of that channel when assisted by free, forward classical communication?

The above assumptions are useful for finding simple ways to make protocols dual under resource reversal, and we will exploit them later in our proofs of various capacity theorems in quantum Shannon theory. But it turns out that there is a more clever way to make teleportation and super-dense coding dual under resource reversal. In this chapter, we introduce a new resource—the *noiseless coherent bit channel*. This resource produces "coherent" versions of the teleportation and super-dense coding protocols that are dual under resource reversal. The payoff of this coherent communication technique is that we can exploit it to simplify the proofs of various coding theorems of quantum Shannon theory. It also leads to a deeper understanding of the relationship between the teleportation and dense coding protocols from the previous chapter.

7.1 Definition of Coherent Communication

We begin by introducing the coherent bit channel as a classical channel that has quantum feedback. Recall from Exercise 6.1.1 that a classical bit channel is equivalent to a dephasing

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

channel that dephases in the computational basis with dephasing parameter $p = 1/2$. The CPTP map corresponding to the dephasing channel is as follows:

$$\mathcal{N}(\rho) = \frac{1}{2}(\rho + Z\rho Z). \tag{7.7}$$

The isometric extension $U_{\mathcal{N}}$ of the above channel then follows by applying (5.25):

$$U_{\mathcal{N}} = \frac{1}{\sqrt{2}}\Big(I^{A\to B} \otimes |+\rangle^E + Z^{A\to B} \otimes |-\rangle^E\Big), \tag{7.8}$$

where we choose the orthonormal basis states of the environment E to be $|+\rangle$ and $|-\rangle$ (recall that we have unitary freedom in the choice of the basis states for the environment). It is straightforward to show that the isometry $U_{\mathcal{N}}$ is as follows by expanding the operators I and Z and the states $|+\rangle$ and $|-\rangle$:

$$U_{\mathcal{N}} = |0\rangle^B\langle 0|^A \otimes |0\rangle^E + |1\rangle^B\langle 1|^A \otimes |1\rangle^E. \tag{7.9}$$

Thus, a classical bit channel is equivalent to the following linear map:

$$|i\rangle^A \to |i\rangle^B|i\rangle^E : i \in \{0, 1\}. \tag{7.10}$$

A coherent bit channel is similar to the above classical bit channel map, with the exception that Alice regains control of the environment of the channel:

$$|i\rangle^A \to |i\rangle^B|i\rangle^A : i \in \{0, 1\}. \tag{7.11}$$

In this sense, the coherent channel is a quantum feedback channel. "Coherence" in this context is also synonymous with linearity—the maintenance and linear transformation of superposed states. The coherent bit channel is similar to classical copying because it copies the basis states while maintaining coherent superpositions. We denote the resource of a coherent bit channel as follows:

$$[q \to qq]. \tag{7.12}$$

Figure 7.1 provides a visual depiction of the coherent bit channel.

Exercise 7.1.1 Show that the following resource inequality holds:

$$[q \to qq] \geq [c \to c]. \tag{7.13}$$

That is, devise a protocol that generates a noiseless classical bit channel with one use of a noiseless coherent bit channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

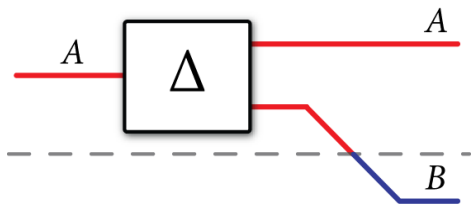

Figure 7.1: The above figure depicts the operation of a coherent bit channel. It is the "coherification" of a classical bit channel in which the sender A has access to the environment's output. For this reason, we also refer to it as the *quantum feedback channel*.

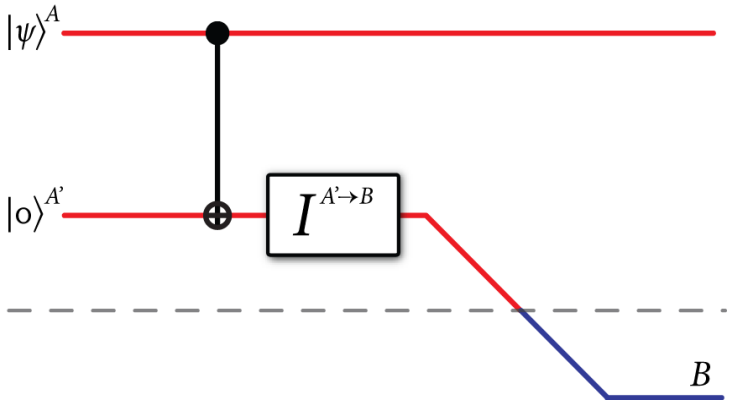

Figure 7.2: A simple protocol to implement a noiseless coherent channel with one use of a noiseless qubit channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

7.2 Implementations of a Coherent Bit Channel

How might we actually implement a coherent bit channel? The simplest way to do so is with the aid of a local CNOT gate and a noiseless qubit channel. The protocol proceeds as follows (Figure 7.2 illustrates the protocol):

1. Alice possesses an information qubit in the state $|\psi\rangle^A \equiv \alpha|0\rangle^A + \beta|1\rangle^A$. She prepares an ancilla qubit in the state $|0\rangle^{A'}$.

2. Alice performs a local CNOT gate from qubit A to qubit A'. The resulting state is

$$\alpha|0\rangle^A|0\rangle^{A'} + \beta|1\rangle^A|1\rangle^{A'}. \tag{7.14}$$

3. Alice transmits qubit A' to Bob with one use of a noiseless qubit channel $I^{A'\to B}$. The resulting state is

$$\alpha|0\rangle^A|0\rangle^B + \beta|1\rangle^A|1\rangle^B, \tag{7.15}$$

and it is now clear that Alice and Bob have implemented a noiseless coherent bit channel as defined in (7.11).

The above protocol implements the following resource inequality:

$$[q \to q] \geq [q \to qq], \tag{7.16}$$

demonstrating that quantum communication generates coherent communication.

Exercise 7.2.1 Show that the following resource inequality holds:

$$[q \to qq] \geq [qq]. \tag{7.17}$$

That is, devise a protocol that generates a noiseless ebit with one use of a noiseless coherent bit channel.

Exercise 7.2.2 Show that the following two resource inequalities cannot hold.

$$[q \to qq] \geq [q \to q], \tag{7.18}$$
$$[qq] \geq [q \to qq]. \tag{7.19}$$

We now have the following chain of resource inequalities:

$$[q \to q] \geq [q \to qq] \geq [qq]. \tag{7.20}$$

Thus, the power of the coherent bit channel lies in between that of a noiseless qubit channel and a noiseless ebit.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 7.2.3 Another way to implement a noiseless coherent bit channel is with a variation of teleportation that we name *coherent communication assisted by entanglement and classical communication*. Suppose that Alice and Bob share an ebit $|\Phi^+\rangle^{AB}$. Alice can append an ancilla qubit $|0\rangle^{A'}$ to this state, perform a local CNOT from A to A' to give the following state:

$$|\Phi_{\text{GHZ}}\rangle^{AA'B} = \frac{1}{\sqrt{2}}\Big(|000\rangle^{AA'B} + |111\rangle^{AA'B}\Big). \tag{7.21}$$

Alice prepends an information qubit $|\psi\rangle^{A_1} \equiv \alpha|0\rangle^{A_1} + \beta|1\rangle^{A_1}$ to the above state so that the global state is as follows:

$$|\psi\rangle^{A_1}|\Phi_{\text{GHZ}}\rangle^{AA'B}. \tag{7.22}$$

Suppose Alice performs the usual teleportation operations on systems A_1, A, and A'. Give the steps that Alice and Bob should perform in order to generate the state $\alpha|0\rangle^{A'}|0\rangle^{B}+\beta|1\rangle^{A'}|1\rangle^{B}$, thus implementing a noiseless coherent bit channel. *Hint:* The resource inequality for this protocol is as follows:

$$[qq] + [c \to c] \geq [q \to qq]. \tag{7.23}$$

Exercise 7.2.4 Determine a qudit version of coherent communication assisted by classical communication and entanglement by modifying the steps in the above protocol.

7.3 Coherent Dense Coding

In the previous section, we introduced two protocols that implement a noiseless coherent bit channel: the simple method in the previous section and coherent communication assisted by classical communication and entanglement (Exercise 7.2.3). We now introduce a different method for implementing two coherent bit channels that makes more judicious use of available resources. We name it *coherent super-dense coding* because it is a coherent version of the super-dense coding protocol.

The protocol proceeds as follows (Figure 7.3 depicts the protocol):

1. Alice and Bob share one ebit in the state $|\Phi^+\rangle^{AB}$ before the protocol begins.

2. Alice first prepares two qubits A_1 and A_2 in the state $|a_1\rangle^{A_1}|a_2\rangle^{A_2}$ and prepends this state to the ebit. The global state is as follows:

$$|a_1\rangle^{A_1}|a_2\rangle^{A_2}\big|\Phi^+\big\rangle^{AB}, \tag{7.24}$$

   where a_1 and a_2 are binary-valued. This preparation step is reminiscent of the super-dense coding protocol (recall that, in the super-dense coding protocol, Alice has two classical bits she would like to communicate).

3. Alice performs a CNOT gate from register A_2 to register A and performs a controlled-Z gate from register A_1 to register A. The resulting state is as follows:

$$|a_1\rangle^{A_1}|a_2\rangle^{A_2}(Z^{a_1}X^{a_2})^{A}\big|\Phi^+\big\rangle^{AB}. \tag{7.25}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

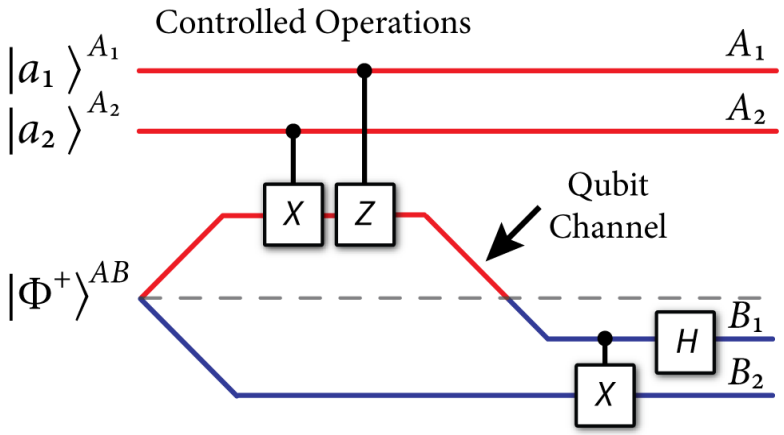

Figure 7.3: The above figure depicts the protocol for coherent super-dense coding.

4. Alice transmits the qubit in register A to Bob. We rename this register as B_1 and Bob's other register B as B_2.

5. Bob performs a CNOT gate from his register B_1 to B_2 and performs a Hadamard gate on B_1. The final state is as follows:

$$|a_1\rangle^{A_1}|a_2\rangle^{A_2}|a_1\rangle^{B_1}|a_2\rangle^{B_2}. \tag{7.26}$$

The above protocol implements two coherent bit channels: one from A_1 to B_1 and another from A_2 to B_2. You can check that the protocol works for arbitrary superpositions of two-qubit states on A_1 and A_2—it is for this reason that this protocol implements two coherent bit channels. The resource inequality corresponding to coherent super-dense coding is

$$[qq] + [q \to q] \geq 2[q \to qq]. \tag{7.27}$$

Exercise 7.3.1 Construct a *qudit* version of coherent super-dense coding that implements the following resource inequality:

$$\log d[qq] + \log d[q \to q] \geq 2 \log d[q \to qq]. \tag{7.28}$$

(*Hint*: The qudit analog of a controlled-NOT gate is

$$\sum_{i=0}^{d-1} |i\rangle\langle i| \otimes X(i), \tag{7.29}$$

where X is defined in (3.209). The qudit analog of the controlled-Z gate is

$$\sum_{j=0}^{d-1} |j\rangle\langle j| \otimes Z(j), \tag{7.30}$$

where Z is defined in (3.212). The qudit analog of the Hadamard gate is the Fourier transform gate.)

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

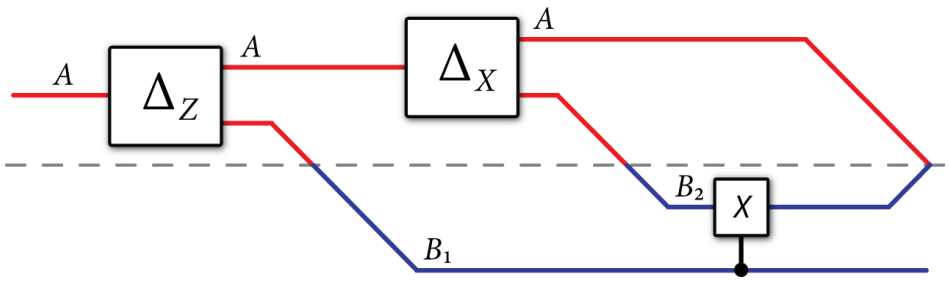

Figure 7.4: The above figure depicts the protocol for coherent teleportation.

7.4 Coherent Teleportation

We now introduce a coherent version of the teleportation protocol that we name *coherent teleportation*. Let a Z coherent bit channel Δ_Z be one that copies eigenstates of the Z operator (this is as we defined a coherent bit channel before). Let an X coherent bit channel Δ_X be one that copies eigenstates of the X operator:

$$\Delta_X : |+\rangle^A \rightarrow |+\rangle^A |+\rangle^B, \tag{7.31}$$

$$|-\rangle^A \rightarrow |-\rangle^A |-\rangle^B. \tag{7.32}$$

It does not really matter which basis we use to define a coherent bit channel—it just matters that it copies the orthogonal states of some basis.

Exercise 7.4.1 Show how to simulate an X coherent bit channel with a Z coherent bit channel and local operations.

The protocol proceeds as follows (Figure 7.4 depicts the protocol):

1. Alice possesses an information qubit $|\psi\rangle^A$ where

$$|\psi\rangle^A \equiv \alpha|0\rangle^A + \beta|1\rangle^A. \tag{7.33}$$

She sends her qubit through a Z coherent bit channel:

$$|\psi\rangle^A \quad \underrightarrow{\Delta_Z} \quad \alpha|0\rangle^A|0\rangle^{B_1} + \beta|1\rangle^A|1\rangle^{B_1} \equiv |\tilde{\psi}\rangle^{AB_1}. \tag{7.34}$$

Let us rewrite the above state $|\tilde{\psi}\rangle^{AB_1}$ as follows:

$$|\tilde{\psi}\rangle^{AB_1} = \alpha\left(\frac{|+\rangle^A + |-\rangle^A}{\sqrt{2}}\right)|0\rangle^{B_1} + \beta\left(\frac{|+\rangle^A - |-\rangle^A}{\sqrt{2}}\right)|1\rangle^{B_1} \tag{7.35}$$

$$= \frac{1}{\sqrt{2}}\left[|+\rangle^A\left(\alpha|0\rangle^{B_1} + \beta|1\rangle^{B_1}\right) + |-\rangle^A\left(\alpha|0\rangle^{B_1} - \beta|1\rangle^{B_1}\right)\right]. \tag{7.36}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

2. Alice sends her qubit A through an X coherent bit channel with output systems A and B_2:

$$|\tilde{\psi}\rangle^{AB_1} \quad \underrightarrow{\Delta_X} \quad \frac{1}{\sqrt{2}}|+\rangle^A|+\rangle^{B_2}\Big(\alpha|0\rangle^{B_1}+\beta|1\rangle^{B_1}\Big) + \frac{1}{\sqrt{2}}|-\rangle^A|-\rangle^{B_2}\Big(\alpha|0\rangle^{B_1}-\beta|1\rangle^{B_1}\Big) \tag{7.37}$$

3. Bob then performs a CNOT gate from qubit B_1 to qubit B_2. Consider that the action of the CNOT gate with the source qubit in the computational basis and the target qubit in the $+/-$ basis is as follows:

$$|0\rangle|+\rangle \rightarrow |0\rangle|+\rangle, \tag{7.38}$$
$$|0\rangle|-\rangle \rightarrow |0\rangle|-\rangle, \tag{7.39}$$
$$|1\rangle|+\rangle \rightarrow |1\rangle|+\rangle, \tag{7.40}$$
$$|1\rangle|-\rangle \rightarrow -|1\rangle|-\rangle, \tag{7.41}$$

so that the last entry catches a phase of π ($e^{i\pi} = -1$). Then this CNOT gate brings the overall state to

$$\frac{1}{\sqrt{2}}\Big[|+\rangle^A|+\rangle^{B_2}\Big(\alpha|0\rangle^{B_1}+\beta|1\rangle^{B_1}\Big) + |-\rangle^A|-\rangle^{B_2}\Big(\alpha|0\rangle^{B_1}+\beta|1\rangle^{B_1}\Big)\Big]$$
$$= \frac{1}{\sqrt{2}}\Big[|+\rangle^A|+\rangle^{B_2}|\psi\rangle^{B_1} + |-\rangle^A|-\rangle^{B_2}|\psi\rangle^{B_1}\Big] \tag{7.42}$$
$$= \frac{1}{\sqrt{2}}\Big[|+\rangle^A|+\rangle^{B_2} + |-\rangle^A|-\rangle^{B_2}\Big]|\psi\rangle^{B_1} \tag{7.43}$$
$$= \left|\Phi^+\right\rangle^{AB_2}|\psi\rangle^{B_1} \tag{7.44}$$

Thus, Alice teleports her information qubit to Bob, and both Alice and Bob possess one ebit at the end of the protocol.

The resource inequality for coherent teleportation is as follows:

$$2[q \rightarrow qq] \geq [qq] + [q \rightarrow q]. \tag{7.45}$$

Exercise 7.4.2 Show how a cobit channel and an ebit can generate a GHZ state. That is, demonstrate a protocol that implements the following resource inequality:

$$[qq]_{AB} + [q \rightarrow qq]_{BC} \geq [qqq]_{ABC}. \tag{7.46}$$

Exercise 7.4.3 Outline the qudit version of the above coherent teleportation protocol. The protocol should implement the following resource inequality:

$$2\log d[q \rightarrow qq] \geq \log d[qq] + \log d[q \rightarrow q]. \tag{7.47}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 7.4.4 Outline a catalytic version of the coherent teleportation protocol by modifying the original teleportation protocol. Let Alice possess an information qubit $|\psi\rangle^{A'}$ and let Alice and Bob share an ebit $|\Phi^+\rangle^{AB}$. Replace the Bell measurement with a controlled-NOT and Hadamard gate, replace the classical bit channels with coherent bit channels, and replace Bob's conditional unitary operations with controlled unitary operations. The resulting resource inequality should be of the form:

$$2[q \rightarrow qq] + [qq] \geq [q \rightarrow q] + 2[qq]. \tag{7.48}$$

This protocol is catalytic in the sense that it gives the resource inequality in (7.45) when we cancel one ebit from each side.

7.5 The Coherent Communication Identity

The fundamental result of this chapter is the *coherent communication identity*:

$$2[q \rightarrow qq] = [qq] + [q \rightarrow q]. \tag{7.49}$$

We obtain this identity by combining the resource inequality for coherent super-dense coding in (7.27) and the resource inequality for coherent teleportation in (7.45). The coherent communication identity demonstrates that coherent super-dense coding and coherent teleportation are dual under resource reversal—the resources that coherent teleportation consumes are the same as those that coherent super-dense coding generates and vice versa.

The major application of the coherent communication identity is in noisy quantum Shannon theory. We will find later that its application is in the "upgrading" of protocols that output private classical information. Suppose that a protocol outputs private classical bits. The super-dense coding protocol is one such example, as the last paragraph of Section 6.2.3 argues. Then it is possible to upgrade the protocol by making it coherent, similar to the way in which we made super-dense coding coherent by replacing conditional unitary operations with controlled unitary operations.

We make this idea more precise with an example. The resource inequality for entanglement-assisted classical coding (discussed in more detail in Chapter 20) has the following form:

$$\langle \mathcal{N} \rangle + E[qq] \geq C[c \rightarrow c], \tag{7.50}$$

where $\mathcal{N}$ is a noisy quantum channel that connects Alice to Bob, E is some rate of entanglement consumption, and C is some rate of classical communication. It is possible to upgrade the generated classical bits to coherent bits, for reasons that are similar to those that we used in the upgrading of super-dense coding. The resulting resource inequality has the following form:

$$\langle \mathcal{N} \rangle + E[qq] \geq C[q \rightarrow qq]. \tag{7.51}$$

We can now employ the coherent communication identity in (7.49) and argue that any protocol that implements the above resource inequality can implement the following one:

$$\langle \mathcal{N} \rangle + E[qq] \geq \frac{C}{2}[q \rightarrow q] + \frac{C}{2}[qq], \tag{7.52}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

merely by using the generated coherent bits in a coherent super-dense coding protocol. We can then make a "catalytic argument" to cancel the ebits on both sides of the resource inequality. The final resource inequality is as follows:

$$\langle \mathcal{N} \rangle + \left(E - \frac{C}{2} \right) [qq] \geq \frac{C}{2} [q \rightarrow q]. \tag{7.53}$$

The above resource inequality corresponds to a protocol for *entanglement-assisted quantum coding* (also known as the *father protocol*), and it turns out to be optimal for some channels as this protocol's converse theorem shows. This optimality is due to the efficient translation of classical bits to coherent bits and the application of the coherent communication identity.

7.6 History and Further Reading

Harrow introduced the idea of coherent communication in Ref. [123]. Later, some researchers studied the continuous-variable coherent channel as well [254]. The idea of coherent communication has many applications in quantum Shannon theory which we will study in later chapters.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 8

The Unit Resource Capacity Region

In Chapter 6, we presented the three unit protocols of teleportation, super-dense coding, and entanglement distribution. The physical arguments in Section 6.3 prove that each of these protocols are individually optimal. For example, recall that the entanglement distribution protocol is optimal because two parties cannot generate more than one ebit from the use of one noiseless qubit channel.

In this chapter, we show that these three protocols are actually the most important protocols—we do not need to consider any other protocols when the noiseless resources of classical communication, quantum communication, and entanglement are available. Combining these three protocols together is the best that one can do with the unit resources.

In this sense, this chapter gives the first example in this book of a converse proof of a capacity theorem. We construct a three-dimensional region, known as the unit resource achievable region, that the three unit protocols fill out. The converse proof of this chapter employs good physical arguments to show that the unit resource achievable region is optimal, and we can then refer to it as the unit resource capacity region. We later exploit the development here when we get to the study of trade-off capacities (see Chapter 24).

8.1 The Unit Resource Achievable Region

Let us first recall the resource inequalities for the three unit protocols. The resource inequality for teleportation is

$$2[c \rightarrow c] + [qq] \geq [q \rightarrow q], \tag{8.1}$$

while that for super-dense coding is

$$[q \rightarrow q] + [qq] \geq 2[c \rightarrow c], \tag{8.2}$$

and that for entanglement distribution is as follows:

$$[q \rightarrow q] \geq [qq]. \tag{8.3}$$

Each of the resources $[q \rightarrow q]$, $[qq]$, $[c \rightarrow c]$ is a *unit resource*.

The above three unit protocols are sufficient to recover all other unit protocols. For example, we can combine super-dense coding and entanglement distribution to produce the following resource inequality:

$$2[q \to q] + [qq] \geq 2[c \to c] + [qq]. \tag{8.4}$$

The above resource inequality is equivalent to the following one

$$[q \to q] \geq [c \to c], \tag{8.5}$$

after removing the entanglement from both sides and scaling by $1/2$ (we can remove the entanglement here because it acts as a catalytic resource). We can justify this by considering a scenario in which we use the above protocol N times. For the first run of the protocol, we require one ebit to get it started, but then every other run both consumes and generates one ebit, giving

$$2N[q \to q] + [qq] \geq 2N[c \to c] + [qq]. \tag{8.6}$$

Dividing by N gives the rate of the task, and as N becomes large, the use of the initial ebit is negligible. We refer to (8.5) as "classical coding over a noiseless qubit channel."

We can think of the above resource inequalities in a different way. Let us consider a three-dimensional space with points of the form (C, Q, E), where C corresponds to noiseless classical communication, Q corresponds to noiseless quantum communication, and E corresponds to noiseless entanglement. Each point in this space corresponds to a protocol involving the unit resources. A coordinate of a point is negative if the point's corresponding resource inequality consumes that coordinate's corresponding resource, and a coordinate of a point is positive if the point's corresponding resource inequality generates that coordinate's corresponding resource.

For example, the point corresponding to the teleportation protocol is

$$x_{\text{TP}} \equiv (-2, 1, -1), \tag{8.7}$$

because teleportation consumes two noiseless classical bit channels and one ebit to generate one noiseless qubit channel. For similar reasons, the respective points corresponding to super-dense coding and entanglement distribution are as follows:

$$x_{\text{SD}} \equiv (2, -1, -1), \tag{8.8}$$

$$x_{\text{ED}} \equiv (0, -1, 1). \tag{8.9}$$

Figure 8.1 plots these three points in the three-dimensional space of classical communication, quantum communication, and entanglement.

We can execute any of the three unit protocols just one time, or we can execute any one of them m times where m is some positive integer. Executing a protocol m times then gives other points in the three dimensional space. That is, we can also achieve the points mx_{TP}, mx_{SD}, and mx_{ED} for any positive m. This method allows us to fill up a certain portion of the three-dimensional space. Let us also suppose that we can achieve real number amounts

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

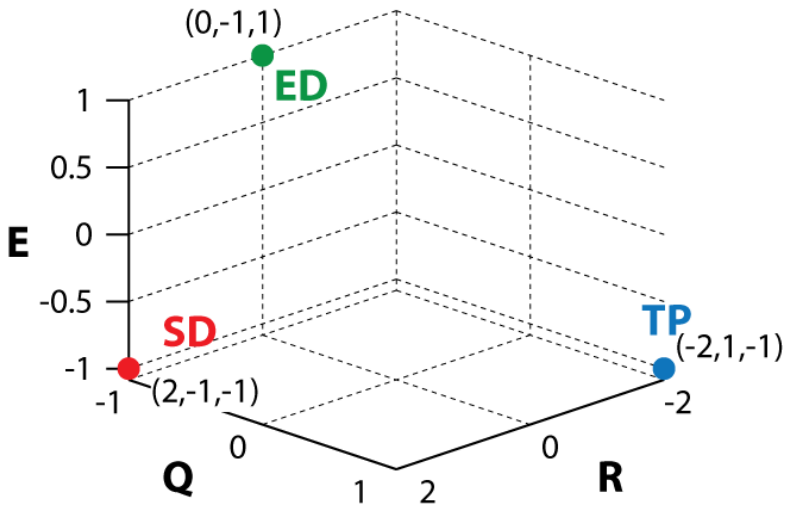

Figure 8.1: The three points corresponding to the three respective unit protocols of entanglement distribution (ED), teleportation (TP), and super-dense coding (SD).

of each protocol. This becomes important later on when we consider combining the three unit protocols in order to achieve certain rates of transmission (a communication rate can be any real number). Thus, we can combine the protocols together to achieve any point of the following form:

$$\alpha x_{\text{TP}} + \beta x_{\text{SD}} + \gamma x_{\text{ED}}, \tag{8.10}$$

where $\alpha, \beta, \gamma \geq 0$.

Let us further establish some notation. Let L denote a line, Q a quadrant, and O an octant in the three-dimensional space (it should be clear from context whether Q refers to quantum communication or "quadrant"). For example, L^{-00} denotes a line going in the direction of negative classical communication:

$$L^{-00} \equiv \{\alpha(-1,0,0) : \alpha \geq 0\}. \tag{8.11}$$

Q^{0+-} denotes the quadrant where there is zero classical communication, generation of quantum communication, and consumption of entanglement:

$$Q^{0+-} \equiv \{\alpha(0,1,0) + \beta(0,0,-1) : \alpha, \beta \geq 0\}. \tag{8.12}$$

O^{+-+} denotes the octant where there is generation of classical communication, consumption of quantum communication, and generation of entanglement:

$$O^{+-+} \equiv \left\{ \begin{array}{c} \alpha(1,0,0) + \beta(0,-1,0) + \gamma(0,0,1) \\ : \alpha, \beta, \gamma \geq 0 \end{array} \right\}. \tag{8.13}$$

It proves useful to have a "set addition" operation between two regions A and B (known as the Minkowski sum):

$$A + B \equiv \{a + b : a \in A, b \in B\}. \tag{8.14}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The following relations hold

$$Q^{0+-} = L^{0+0} + L^{00-}, \tag{8.15}$$
$$O^{+-+} = L^{+00} + L^{0-0} + L^{00+}, \tag{8.16}$$

by using the above definition.

The following geometric objects lie in the (C, Q, E) space:

1. The "line of teleportation" L_{TP} is the following set of points:

$$L_{\text{TP}} \equiv \{\alpha(-2, 1, -1) : \alpha \geq 0\}. \tag{8.17}$$

2. The "line of super-dense coding" L_{SD} is the following set of points:

$$L_{\text{SD}} \equiv \{\beta(2, -1, -1) : \beta \geq 0\}. \tag{8.18}$$

3. The "line of entanglement distribution" L_{ED} is the following set of points:

$$L_{\text{ED}} \equiv \{\gamma(0, -1, 1) : \gamma \geq 0\}. \tag{8.19}$$

Definition 8.1.1. *Let $\widetilde{C}_{\text{U}}$ denote the unit resource achievable region. It consists of all linear combinations of the above protocols:*

$$\widetilde{C}_{\text{U}} \equiv L_{\text{TP}} + L_{\text{SD}} + L_{\text{ED}}. \tag{8.20}$$

The following matrix equation gives all achievable triples (C, Q, E) in $\widetilde{C}_{\text{U}}$:

$$\begin{bmatrix} C \\ Q \\ E \end{bmatrix} = \begin{bmatrix} -2 & 2 & 0 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}, \tag{8.21}$$

where $\alpha, \beta, \gamma \geq 0$. We can rewrite the above equation with its matrix inverse:

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} -1/2 & -1/2 & -1/2 \\ 0 & -1/2 & -1/2 \\ -1/2 & -1 & 0 \end{bmatrix} \begin{bmatrix} C \\ Q \\ E \end{bmatrix}, \tag{8.22}$$

in order to express the coefficients α, β, and γ as a function of the rate triples (C, Q, E). The restriction of non-negativity of α, β, and γ gives the following restriction on the achievable rate triples (C, Q, E):

$$C + Q + E \leq 0, \tag{8.23}$$
$$Q + E \leq 0, \tag{8.24}$$
$$C + 2Q \leq 0. \tag{8.25}$$

The above result implies that the achievable region $\widetilde{C}_{\text{U}}$ in (8.20) is equivalent to all rate triples satisfying (8.23-8.25). Figure 8.2 displays the full unit resource achievable region.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

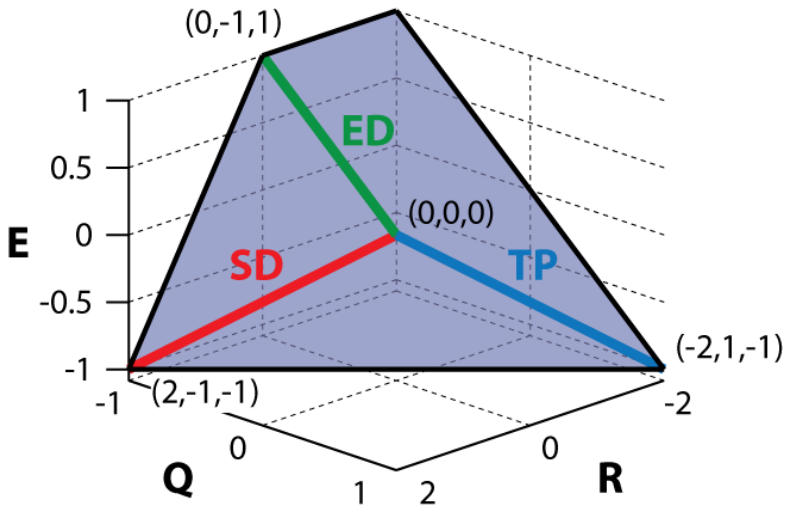

Figure 8.2: The above figure depicts the unit resource achievable region $\widetilde{C}_{\mathrm{U}}$.

Definition 8.1.2. *The unit resource capacity region C_{U} is the closure of the set of all points (C, Q, E) in the C, Q, E space, satisfying the following resource inequality:*

$$0 \geq C[c \to c] + Q[q \to q] + E[qq]. \tag{8.26}$$

The definition states that the unit resource capacity region consists of all those points (C, Q, E) that have corresponding protocols that can implement them. The notation in the above definition may seem slightly confusing at first glance until we recall that a resource with a negative rate implicitly belongs on the left-hand side of the resource inequality.

Theorem 8.1.1 below gives the optimal three-dimensional capacity region for the three unit resources.

Theorem 8.1.1. *The unit resource capacity region C_{U} is equivalent to the unit resource achievable region $\widetilde{C}_{\mathrm{U}}$:*

$$C_{\mathrm{U}} = \widetilde{C}_{\mathrm{U}}. \tag{8.27}$$

Proving the above theorem involves two steps: the *direct coding theorem* and the *converse theorem*. For this case, the *direct coding theorem* establishes that the achievable region $\widetilde{C}_{\mathrm{U}}$ is in the capacity region C_{U}:

$$\widetilde{C}_{\mathrm{U}} \subseteq C_{\mathrm{U}}. \tag{8.28}$$

The *converse theorem*, on the other hand, establishes that the achievable region $\widetilde{C}_{\mathrm{U}}$ is optimal:

$$C_{\mathrm{U}} \subseteq \widetilde{C}_{\mathrm{U}}. \tag{8.29}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

8.2 The Direct Coding Theorem

The result of the direct coding theorem, that $\widetilde{C}_{\mathrm{U}} \subseteq C_{\mathrm{U}}$, is immediate from the definition in (8.20) of the unit resource achievable region $\widetilde{C}_{\mathrm{U}}$, the definition in (8.26) of the unit resource capacity region C_{U}, and the theory of resource inequalities. We can achieve points in the unit resource capacity region simply by considering positive linear combinations of the three unit protocols. The next section shows that the unit resource capacity region consists of all and only those points in the unit resource achievable region.

8.3 The Converse Theorem

We employ the definition of $\widetilde{C}_{\mathrm{U}}$ in (8.20) and consider the eight octants of the (C, Q, E) space individually in order to prove the converse theorem (that $C_{\mathrm{U}} \subseteq \widetilde{C}_{\mathrm{U}}$). Let $(\pm, \pm, \pm)$ denote labels for the eight different octants.

It is possible to demonstrate the optimality of each of these three protocols individually with a contradiction argument as we saw in Chapter 6. However, in the converse proof of Theorem 8.1.1, we show that a mixed strategy combining these three unit protocols is optimal.

We accept the following two postulates and exploit them in order to prove the converse:

1. Entanglement alone cannot generate classical communication or quantum communication or both.

2. Classical communication alone cannot generate entanglement or quantum communication or both.

$(+,+,+)$. This octant of C_{U} is empty because a sender and receiver require some resources to implement classical communication, quantum communication, and entanglement. (They cannot generate a unit resource from nothing!)

$(+,+,-)$. This octant of C_{U} is empty because entanglement alone cannot generate either classical communication or quantum communication or both.

$(+,-,+)$. The task for this octant is to generate a noiseless classical channel of C bits and E ebits of entanglement by consuming $|Q|$ qubits of quantum communication. We thus consider all points of the form (C, Q, E) where $C \geq 0$, $Q \leq 0$, and $E \geq 0$. It suffices to prove the following inequality:

$$C + E \leq |Q|, \tag{8.30}$$

because combining (8.30) with $C \geq 0$ and $E \geq 0$ implies (8.23-8.25). The achievability of $(C, -|Q|, E)$ implies the achievability of the point $(C + 2E, -|Q| - E, 0)$, because we can consume all of the entanglement with super-dense coding (8.2):

$$(C + 2E, -|Q| - E, 0) = (C, -|Q|, E) + (2E, -E, -E). \tag{8.31}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

This new point implies that there is a protocol that consumes $|Q|+E$ noiseless qubit channels to send $C + 2E$ classical bits. The following bound then applies

$$C + 2E \leq |Q| + E, \tag{8.32}$$

because the Holevo bound (Exercise 4.2.2 gives a simpler statement of this bound) states that we can send only one classical bit per qubit. The bound in (8.30) then follows.

$(+,-,-)$. The task for this octant is to simulate a classical channel of size C bits using $|Q|$ qubits of quantum communication and $|E|$ ebits of entanglement. We consider all points of the form (C, Q, E) where $C \geq 0$, $Q \leq 0$, and $E \leq 0$. It suffices to prove the following inequalities:

$$C \leq 2|Q|, \tag{8.33}$$
$$C \leq |Q| + |E|, \tag{8.34}$$

because combining (8.33-8.34) with $C \geq 0$ implies (8.23-8.25). The achievability of $(C, -|Q|, -|E|)$ implies the achievability of $(0, -|Q| + C/2, -|E| - C/2)$, because we can consume all of the classical communication with teleportation (8.1):

$$(0, -|Q| + C/2, -|E| - C/2) = (C, -|Q|, -|E|) + (-C, C/2, -C/2). \tag{8.35}$$

The following bound applies (quantum communication cannot be positive)

$$-|Q| + C/2 \leq 0, \tag{8.36}$$

because entanglement alone cannot generate quantum communication. The bound in (8.33) then follows from the above bound. The achievability of $(C, -|Q|, -|E|)$ implies the achievability of $(C, -|Q| - |E|, 0)$ because we can consume an extra $|E|$ qubit channels with entanglement distribution (8.3):

$$(C, -|Q| - |E|, 0) = (C, -|Q|, -|E|) + (0, -|E|, |E|). \tag{8.37}$$

The bound in (8.34) then applies by the same Holevo bound argument as in the previous octant.

$(-,+,+)$. This octant of C_{U} is empty because classical communication alone cannot generate either quantum communication or entanglement or both.

$(-,+,-)$. The task for this octant is to simulate a quantum channel of size Q qubits using $|E|$ ebits of entanglement and $|C|$ bits of classical communication. We consider all points of the form (C, Q, E) where $C \leq 0$, $Q \geq 0$, and $E \leq 0$. It suffices to prove the following inequalities:

$$Q \leq |E|, \tag{8.38}$$
$$2Q \leq |C|, \tag{8.39}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

because combining them with $C \leq 0$ implies (8.23-8.25). The achievability of the point $(-|C|, Q, -|E|)$ implies the achievability of the point $(-|C|, 0, Q - |E|)$, because we can consume all of the quantum communication for entanglement distribution (8.3):

$$(-|C|, 0, Q - |E|) = (-|C|, Q, -|E|) + (0, -Q, Q). \tag{8.40}$$

The following bound applies (entanglement cannot be positive)

$$Q - |E| \leq 0, \tag{8.41}$$

because classical communication alone cannot generate entanglement. The bound in (8.38) follows from the above bound. The achievability of the point $(-|C|, Q, -|E|)$ implies the achievability of the point $(-|C| + 2Q, 0, -Q - |E|)$, because we can consume all of the quantum communication for super-dense coding (8.2):

$$(-|C| + 2Q, 0, -Q - |E|) = (-|C|, Q, -|E|) + (2Q, -Q, -Q). \tag{8.42}$$

The following bound applies (classical communication cannot be positive)

$$-|C| + 2Q \leq 0, \tag{8.43}$$

because entanglement alone cannot create classical communication. The bound in (8.39) follows from the above bound.

$\mathbf{(-,-,+)}$. The task for this octant is to create E ebits of entanglement using $|Q|$ qubits of quantum communication and $|C|$ bits of classical communication. We consider all points of the form (C, Q, E) where $C \leq 0$, $Q \leq 0$, and $E \geq 0$. It suffices to prove the following inequality:

$$E \leq |Q|, \tag{8.44}$$

because combining it with $Q \leq 0$ and $C \leq 0$ implies (8.23-8.25). The achievability of $(-|C|, -|Q|, E)$ implies the achievability of $(-|C| - 2E, -|Q| + E, 0)$, because we can consume all of the entanglement with teleportation (8.1):

$$(-|C| - 2E, -|Q| + E, 0) = (-|C|, -|Q|, E) + (-2E, E, -E). \tag{8.45}$$

The following bound applies (quantum communication cannot be positive)

$$-|Q| + E \leq 0, \tag{8.46}$$

because classical communication alone cannot generate quantum communication. The bound in (8.44) follows from the above bound.

$\mathbf{(-,-,-)}$. $\widetilde{C}_{\mathrm{U}}$ completely contains this octant.

We have now proved that the set of inequalities in (8.23-8.25) holds for all octants of the (C, Q, E) space. The next exercises ask you to consider similar unit resource achievable regions.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 8.3.1 Consider the resources of public classical communication:

$$[c \to c]_{\text{pub}}, \tag{8.47}$$

private classical communication:

$$[c \to c]_{\text{priv}}, \tag{8.48}$$

and shared secret key:

$$[cc]_{\text{priv}}. \tag{8.49}$$

Public classical communication is equivalent to the following channel:

$$\rho \to \sum_i \langle i|\rho|i\rangle \ |i\rangle\langle i|^B \otimes \sigma_i^E, \tag{8.50}$$

so that an eavesdropper Eve obtains some correlations with the transmitted state ρ. Private classical communication is equivalent to the following channel:

$$\rho \to \sum_i \langle i|\rho|i\rangle \ |i\rangle\langle i|^B \otimes \sigma^E, \tag{8.51}$$

so that Eve's state is independent of the information that Bob receives. Finally, a secret key is a state of the following form:

$$\overline{\Phi}^{AB} \otimes \sigma^E \equiv \left(\frac{1}{d} \sum_i |i\rangle\langle i|^A \otimes |i\rangle\langle i|^B \right) \otimes \sigma^E, \tag{8.52}$$

so that Alice and Bob share maximal classical correlation and Eve's state is independent of it. There are three protocols that relate these three classical resources. Secret key distribution is a protocol that consumes a noiseless private channel to generate a noiseless secret key. It has the following resource inequality:

$$[c \to c]_{\text{priv}} \geq [cc]_{\text{priv}}. \tag{8.53}$$

The one-time pad protocol exploits a shared secret key and a noiseless public channel to generate a noiseless private channel (it simply XORs a bit of secret key with the bit that the sender wants to transmit and this protocol is provably unbreakable if the secret key is perfectly secret). It has the following resource inequality:

$$[c \to c]_{\text{pub}} + [cc]_{\text{priv}} \geq [c \to c]_{\text{priv}}. \tag{8.54}$$

Finally, private classical communication can simulate public classical communication if we assume that Bob has a local register where he can place information and he then gives this to Eve. It has the following resource inequality:

$$[c \to c]_{\text{priv}} \geq [c \to c]_{\text{pub}}. \tag{8.55}$$

Show that these three protocols fill out an optimal achievable region in the space of public classical communication, private classical communication, and secret key. Use the following two postulates to prove optimality: (1) public classical communication alone cannot generate secret key or private classical communication, (2) private key alone cannot generate public or private classical communication.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 8.3.2 Consider the resource of coherent communication from Chapter 7:

$$[q \to qq]. \tag{8.56}$$

Recall the coherent communication identity in (7.49):

$$2[q \to qq] = [q \to q] + [qq]. \tag{8.57}$$

Recall the other resource inequalities for coherent communication:

$$[q \to q] \geq [q \to qq] \geq [qq]. \tag{8.58}$$

Consider a space of points (C, Q, E) where C corresponds to coherent communication, Q to quantum communication, and E to entanglement. Determine the achievable region one obtains with the above resource inequalities and another trivial resource inequality:

$$[qq] \geq 0. \tag{8.59}$$

We interpret the above resource inequality as "entanglement consumption," where Alice simply throws away entanglement.

8.4 History and Further Reading

The unit resource capacity region first appeared in Ref. [160] in the context of trade-off coding. The private unit resource capacity region later appeared in Ref. [252].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Part IV

Tools of Quantum Shannon Theory

CHAPTER 9

Distance Measures

We discussed the major noiseless quantum communication protocols such as teleportation, super-dense coding, their coherent versions, and entanglement distribution in detail in Chapters 6, 7, and 8. Each of these protocols relies on the assumption that noiseless resources are available. For example, the entanglement distribution protocol assumes that a noiseless qubit channel is available to generate a noiseless ebit. This idealization allowed us to develop the main principles of the protocols without having to think about more complicated issues, but in practice, the protocols do not work as expected under the presence of noise.

Given that quantum systems suffer noise in practice, we would like to have a way to determine how well a protocol is performing. The simplest way to do so is to compare the output of an ideal protocol to the output of the actual protocol using a *distance measure* of the two respective output quantum states. That is, suppose that a quantum information processing protocol should ideally output some quantum state $|\psi\rangle$, but the actual output of the protocol is a quantum state with density operator ρ. Then a performance measure $P(|\psi\rangle, \rho)$ should indicate how close the ideal output is to the actual output. Figure 9.1 depicts the comparison of an ideal protocol with another protocol that is noisy.

This chapter introduces two distance measures that allow us to determine how close two quantum states are to each other. The first distance measure that we discuss is the *trace distance* and the second is the *fidelity*. (Though, note that the fidelity is not a distance measure in the strict mathematical sense—nevertheless, we exploit it as a "closeness" mea-

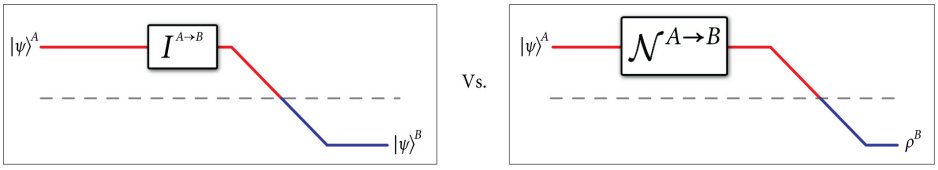

Figure 9.1: A distance measure quantifies how far the output of a given ideal protocol (depicted on the left) is from an actual protocol that exploits a noisy resource (depicted as the noisy quantum channel $\mathcal{N}^{A\to B}$ on the right).

sure of quantum states because it admits an intuitive operational interpretation.) These two measures are mostly interchangeable, but we introduce both because it is often times more convenient in a given situation to use one or the other.

Distance measures are particularly important in quantum Shannon theory because they provide a way for us to determine how well a protocol is performing. Recall that Shannon's method (outlined in Chapter 2) for both the noiseless and noisy coding theorem is to allow for a slight error in a protocol, but to show that this error vanishes in the limit of large block length. In later chapters where we prove quantum coding theorems, we borrow this technique of demonstrating asymptotically small error, with either the trace distance or the fidelity as the measure of performance.

9.1 Trace Distance

We first introduce the trace distance. Our presentation is somewhat mathematical because we exploit norms on linear operators in order to define it. Despite this mathematical flavor, we end this section with an intuitive operational interpretation of the trace distance.

9.1.1 Trace Norm

We begin by defining the *trace norm* or ℓ_1*-norm* $\|M\|_1$ of an Hermitian operator[1] M:

$$\|M\|_1 \equiv \text{Tr}\left\{\sqrt{M^\dagger M}\right\}. \tag{9.1}$$

Recall that any function f applied to an Hermitian operator A is as follows:

$$f(A) \equiv \sum_i f(\alpha_i)|i\rangle\langle i|, \tag{9.2}$$

where $\sum_i \alpha_i |i\rangle\langle i|$ is the spectral decomposition of A. With these two definitions, it is straightforward to show that the trace norm of M is the absolute sum of its eigenvalues:

$$\|M\|_1 = \sum_i |\mu_i|, \tag{9.3}$$

where the spectral decomposition of M is $\sum_i \mu_i |i\rangle\langle i|$.

The trace norm is indeed a *norm* because it satisfies the following three properties: positive definiteness, homogeneity, and the triangle inequality.

Property 9.1.1 (Positive Definiteness) The trace norm of an Hermitian operator M is always positive definite:

$$\|M\|_1 \geq 0. \tag{9.4}$$

The trace norm is null if and only if the operator M is null:

$$\|M\|_1 = 0 \quad \Leftrightarrow \quad M = 0. \tag{9.5}$$

[1]The trace norm applies to any operator, but we restrict ourselves to Hermitian operators to simplify the discussion.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Property 9.1.2 (Homogeneity) For any constant $c \in \mathbb{C}$,

$$\|cM\|_1 = |c|\|M\|_1. \tag{9.6}$$

Property 9.1.3 (Triangle Inequality) For any two operators M and N, the following triangle inequality holds

$$\|M + N\|_1 \leq \|M\|_1 + \|N\|_1. \tag{9.7}$$

Positive definiteness follows because the absolute sum of the eigenvalues of an operator is always non-negative, and the eigenvalues are null (and thus the operator is null) if and only if the absolute sum of the eigenvalues is null. Homogeneity follows because the absolute eigenvalues of cM are equal to $|c|$ times those of M. We later give a proof of the triangle inequality (though, for a special case only).

Two other important properties of the trace norm are its invariance under isometries and convexity. Each of the below properties often arise as useful tools in quantum Shannon theory.

Property 9.1.4 (Isometric invariance) The trace norm is invariant under conjugation by an isometry U:

$$\left\|UMU^\dagger\right\|_1 = \|M\|_1. \tag{9.8}$$

Property 9.1.5 (Convexity) For any two operators M and N and any convex coefficients $\lambda_1, \lambda_2 \geq 0$ such that $\lambda_1 + \lambda_2 = 1$, the following convexity inequality holds

$$\|\lambda_1 M + \lambda_2 N\|_1 \leq \lambda_1\|M\|_1 + \lambda_2\|N\|_1. \tag{9.9}$$

Isometric invariance holds because M and $UMU^\dagger$ have the same eigenvalues. Convexity follows directly from the triangle inequality and homogeneity.

9.1.2 Trace Distance from the Trace Norm

The trace norm induces a natural distance measure between operators, called the *trace distance*.

Definition 9.1.1 (Trace Distance)**.** *Given any two operators M and N, the trace distance between them is as follows:*

$$\|M - N\|_1 = Tr\left\{\sqrt{(M-N)^\dagger(M-N)}\right\}. \tag{9.10}$$

The trace distance is especially useful as a measure of the distinguishability of two quantum states with respective density operators ρ and σ. The following bounds apply to the trace distance between any two density operators ρ and σ:

$$0 \leq \|\rho - \sigma\|_1 \leq 2. \tag{9.11}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The lower bound applies when two quantum states are equivalent—quantum states ρ and σ are equivalent to each other if and only if their trace distance is zero. The physical implication of null trace distance is that no measurement can distinguish ρ from σ. The upper bound follows from the triangle inequality:

$$\|\rho - \sigma\|_1 \leq \|\rho\|_1 + \|\sigma\|_1 = 2. \tag{9.12}$$

The trace distance is maximum when ρ and σ have support on orthogonal subspaces. The physical implication of maximal trace distance is that there exists a measurement that can perfectly distinguish ρ from σ. We discuss these operational interpretations of the trace distance in more detail in Section 9.1.4.

Exercise 9.1.1 Show that the trace distance between two qubit density operators ρ and σ is equal to the Euclidean distance between their respective Bloch vectors $\overrightarrow{r}$ and $\overrightarrow{s}$, where

$$\rho = \frac{1}{2}(I + \overrightarrow{r} \cdot \overrightarrow{\sigma}), \qquad \sigma = \frac{1}{2}(I + \overrightarrow{s} \cdot \overrightarrow{\sigma}). \tag{9.13}$$

That is, show that

$$\|\rho - \sigma\|_1 = \|\overrightarrow{r} - \overrightarrow{s}\|_2. \tag{9.14}$$

Exercise 9.1.2 Show that the trace distance obeys a telescoping property:

$$\|\rho_1 \otimes \rho_2 - \sigma_1 \otimes \sigma_2\|_1 \leq \|\rho_1 - \sigma_1\|_1 + \|\rho_2 - \sigma_2\|_1, \tag{9.15}$$

for any density operators ρ_1, ρ_2, σ_1, σ_2. (Hint: First prove that

$$\|\rho \otimes \omega - \sigma \otimes \omega\|_1 = \|\rho - \sigma\|_1, \tag{9.16}$$

for any density operators ρ, σ, ω.)

Exercise 9.1.3 Show that the trace distance is invariant under an isometric operation U:

$$\|\rho - \sigma\|_1 = \left\|U\rho U^\dagger - U\sigma U^\dagger\right\|_1. \tag{9.17}$$

The physical implication of (9.17) is that an isometry applied to both states does not increase or decrease the distinguishability of the two states.

9.1.3 Trace Distance as a Probability Difference

We now state and prove an important lemma that gives an alternative and useful way for characterizing the trace distance. This particular characterization finds application in many proofs of the lemmas that follow concerning trace distance.

Lemma 9.1.1. *The trace distance $\|\rho - \sigma\|_1$ between quantum states ρ and σ is equal to twice the largest probability difference that two states ρ and σ could give to the same measurement outcome Λ:*

$$\|\rho - \sigma\|_1 = 2 \max_{0 \leq \Lambda \leq I} \text{Tr}\{\Lambda(\rho - \sigma)\}. \tag{9.18}$$

The above maximization is with respect to all positive operators Λ with eigenvalues bounded from above by 1.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Proof. Consider that the difference operator $\rho - \sigma$ is Hermitian and we can diagonalize it as follows:

$$\begin{aligned} \rho - \sigma &= UDU^\dagger && (9.19)\\ &= U\big(D^+ - D^-\big)U^\dagger && (9.20)\\ &= UD^+U^\dagger - UD^-U^\dagger, && (9.21)\end{aligned}$$

where U is a unitary matrix of orthonormal eigenvectors, D is a diagonal matrix of eigenvalues, D^+ is a diagonal matrix whose elements are the positive elements of D, and D^- is a diagonal matrix whose elements are the absolute value of the negative elements of D. We make the following assignments:

$$\begin{aligned} \alpha^+ &\equiv UD^+U^\dagger, && (9.22)\\ \alpha^- &\equiv UD^-U^\dagger. && (9.23)\end{aligned}$$

so that $\rho - \sigma = \alpha^+ - \alpha^-$. Let Π^+ and Π^- be the projectors onto the respective eigenspaces of α^+ and α^-. The projectors obey the orthogonality property $\Pi^+\Pi^- = 0$ because the eigenvectors of $\rho - \sigma$ are orthonormal. So the following properties hold:

$$\begin{aligned} \Pi^+(\rho-\sigma)\Pi^+ &= \Pi^+\big(\alpha^+ - \alpha^-\big)\Pi^+ = \Pi^+\alpha^+\Pi^+ = \alpha^+, && (9.24)\\ \Pi^-(\rho-\sigma)\Pi^- &= \Pi^-\big(\alpha^+ - \alpha^-\big)\Pi^- = -\Pi^-\alpha^-\Pi^- = -\alpha^-. && (9.25)\end{aligned}$$

The following property holds as well

$$|\rho - \sigma| = \big|\alpha^+ - \alpha^-\big| = \alpha^+ + \alpha^- \tag{9.26}$$

because the supports of α^+ and α^- are orthogonal and the absolute value of the operator $\alpha^+ - \alpha^-$ takes the absolute value of its eigenvalues. Therefore,

$$\begin{aligned} \|\rho - \sigma\|_1 &= \mathrm{Tr}\{|\rho - \sigma|\} && (9.27)\\ &= \mathrm{Tr}\big\{\alpha^+ + \alpha^-\big\} && (9.28)\\ &= \mathrm{Tr}\big\{\alpha^+\big\} + \mathrm{Tr}\big\{\alpha^-\big\}. && (9.29)\end{aligned}$$

But

$$\begin{aligned} \mathrm{Tr}\big\{\alpha^+\big\} - \mathrm{Tr}\big\{\alpha^-\big\} &= \mathrm{Tr}\big\{\alpha^+ - \alpha^-\big\} && (9.30)\\ &= \mathrm{Tr}\{\rho - \sigma\} && (9.31)\\ &= \mathrm{Tr}\{\rho\} - \mathrm{Tr}\{\sigma\} && (9.32)\\ &= 0, && (9.33)\end{aligned}$$

where the last equality follows because both quantum states have unit trace. Therefore, $\mathrm{Tr}\{\alpha^+\} = \mathrm{Tr}\{\alpha^-\}$ and

$$\|\rho - \sigma\|_1 = 2\ \mathrm{Tr}\big\{\alpha^+\big\}. \tag{9.34}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Consider then that

$$\begin{aligned} 2\text{Tr}\{\Pi^+(\rho-\sigma)\} &= 2\text{Tr}\{\Pi^+(\alpha^+-\alpha^-)\} && (9.35)\\ &= 2\text{Tr}\{\Pi^+\alpha^+\} && (9.36)\\ &= 2\text{Tr}\{\alpha^+\} && (9.37)\\ &= \|\rho-\sigma\|_1. && (9.38)\end{aligned}$$

Now we prove that the operator Π^+ is the maximizing one. Let Λ be any positive operator with spectrum bounded above by unity. Then

$$\begin{aligned} 2\text{Tr}\{\Lambda(\rho-\sigma)\} &= 2\text{Tr}\{\Lambda(\alpha^+-\alpha^-)\} && (9.39)\\ &\leq 2\text{Tr}\{\Lambda\alpha^+\} && (9.40)\\ &\leq 2\text{Tr}\{\alpha^+\} && (9.41)\\ &= \|\rho-\sigma\|_1. && (9.42)\end{aligned}$$

The first inequality follows because Λ and α^- are positive and thus $\text{Tr}\{\Lambda\alpha^-\}$ is positive. The second inequality holds because $\Lambda \leq I$. The final equality follows from (9.34). □

Exercise 9.1.4 Show that the trace norm of any Hermitian operator ω is given by the following optimization:

$$\|\omega\|_1 = \max_{-I\leq\Lambda\leq I} \text{Tr}\{\Lambda\omega\}. \tag{9.43}$$

9.1.4 Operational Interpretation of the Trace Distance

We now provide an operational interpretation of the trace distance as the distinguishability of two quantum states. The interpretation results from a hypothesis testing scenario. Suppose that Alice prepares one of two quantum states ρ_0 or ρ_1 for Bob to distinguish. Suppose further that it is equally likely *a priori* for her to prepare either ρ_0 or ρ_1. Let X denote the Bernoulli random variable assigned to the prior probabilities so that $p_X(0) = p_X(1) = 1/2$. Bob can perform a binary POVM with elements $\{\Lambda_0, \Lambda_1\}$ to distinguish the two states. That is, Bob guesses the state in question is ρ_0 if he receives outcome "0" from the measurement or he guesses the state in question is ρ_1 if he receives outcome "1" from the measurement. Let Y denote the Bernoulli random variable assigned to the classical outcomes of his measurement. The probability of error p_e for this hypothesis testing scenario is the sum of the probability of detecting "0" when the state is ρ_1 (a so-called Type II error) and the probability of detecting "1" when the state is ρ_0 (a so-called Type I error):

$$\begin{aligned} p_e &= p_{Y|X}(0|1)p_X(1) + p_{Y|X}(1|0)p_X(0) && (9.44)\\ &= \text{Tr}\{\Lambda_0\rho_1\}\frac{1}{2} + \text{Tr}\{\Lambda_1\rho_0\}\frac{1}{2}. && (9.45)\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can simplify this expression using the completeness relation $\Lambda_0 + \Lambda_1 = I$,

$$p_e = \frac{1}{2}(\text{Tr}\{\Lambda_0\rho_1\} + \text{Tr}\{(I - \Lambda_0)\rho_0\}) \tag{9.46}$$

$$= \frac{1}{2}(\text{Tr}\{\Lambda_0\rho_1\} + \text{Tr}\{\rho_0\} - \text{Tr}\{\Lambda_0\rho_0\}) \tag{9.47}$$

$$= \frac{1}{2}(\text{Tr}\{\Lambda_0\rho_1\} + 1 - \text{Tr}\{\Lambda_0\rho_0\}) \tag{9.48}$$

$$= \frac{1}{4}(2 - 2\text{Tr}\{\Lambda_0(\rho_0 - \rho_1)\}). \tag{9.49}$$

Now Bob has freedom in choosing the POVM $\{\Lambda_0, \Lambda_1\}$ to distinguish the states ρ_0 and ρ_1 and he would like to choose one that minimizes the probability of error p_e. Thus, we can rewrite the error probability as follows:

$$p_e = \min_{\Lambda_0,\Lambda_1} \frac{1}{4}(2 - 2\text{Tr}\{\Lambda_0(\rho_0 - \rho_1)\}). \tag{9.50}$$

The minimization problem becomes a maximization as a result of the negative sign:

$$p_e = \frac{1}{4}\left(2 - 2\max_{\Lambda_0,\Lambda_1} \text{Tr}\{\Lambda_0(\rho_0 - \rho_1)\}\right). \tag{9.51}$$

We can rewrite the above quantity in terms of the trace distance using its characterization in Lemma 9.1.1 because the expression inside of the maximization involves only the operator Λ_0:

$$p_e = \frac{1}{4}(2 - \|\rho_0 - \rho_1\|_1). \tag{9.52}$$

Thus, the trace distance has the operational interpretation that it leads to the minimum probability of error in distinguishing two quantum states ρ_0 and ρ_1 in a quantum hypothesis testing experiment. From the above expression for probability of error, it is clear that the states are indistinguishable when $\|\rho_0 - \rho_1\|_1$ is null. That is, it is just as good for Bob to guess randomly what the state might be. On the other hand, the states are perfectly distinguishable when $\|\rho_0 - \rho_1\|_1$ is maximal and the measurement that distinguishes them consists of two projectors: one projects onto the positive eigenspace of $\rho_0 - \rho_1$ and the other projects onto the negative eigenspace of $\rho_0 - \rho_1$.

Exercise 9.1.5 Repeat the above derivation to show that the trace distance admits an operational interpretation in terms of the probability of guessing correctly in quantum hypothesis testing.

Exercise 9.1.6 Suppose that the prior probabilities in the above hypothesis testing scenario are not uniform but are rather equal to p_0 and p_1. Show that the probability of error is instead given by

$$p_e = p_0 - \frac{1}{2}\|p_0\rho_0 - p_1\rho_1\|_1. \tag{9.53}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

9.1.5 Trace Distance Lemmas

We present several useful corollaries of Lemma 9.1.1 and their corresponding proofs. These corollaries include the triangle inequality, measurement on approximate states, and monotonicity of trace distance. Each of these corollaries finds application in many proofs in quantum Shannon theory.

Lemma 9.1.2 (Triangle Inequality)**.** *The trace distance obeys a triangle inequality. For any three quantum states ρ, σ, and τ, the following inequality holds*

$$\|\rho - \sigma\|_1 \leq \|\rho - \tau\|_1 + \|\tau - \sigma\|_1. \tag{9.54}$$

Proof. Pick Π as the maximizing operator for $\|\rho - \sigma\|_1$ (according to Lemma 9.1.1) so that

$$\begin{aligned} \|\rho - \sigma\|_1 &= 2\text{Tr}\{\Pi(\rho - \sigma)\} && (9.55)\\ &= 2\text{Tr}\{\Pi(\rho - \tau)\} + 2\text{Tr}\{\Pi(\tau - \sigma)\} && (9.56)\\ &\leq \|\rho - \tau\|_1 + \|\tau - \sigma\|_1. && (9.57) \end{aligned}$$

The last inequality follows because the operator Π maximizing $\|\rho - \sigma\|_1$ in general is not the same operator that maximizes both $\|\rho - \tau\|_1$ and $\|\tau - \sigma\|_1$. □

Corollary 9.1.1 (Measurement on Approximately Close States)**.** *Suppose we have two quantum states ρ and σ and an operator Π where $0 \leq \Pi \leq I$. Then*

$$\text{Tr}\{\Pi\rho\} \geq \text{Tr}\{\Pi\sigma\} - \|\rho - \sigma\|_1 \tag{9.58}$$

Proof. Consider the following arguments.

$$\begin{aligned} \|\rho - \sigma\|_1 &= 2 \max_{0 \leq \Lambda \leq I} \{\text{Tr}\{\Lambda(\sigma - \rho)\}\} && (9.59)\\ &\geq \max_{0 \leq \Lambda \leq I} \{\text{Tr}\{\Lambda(\sigma - \rho)\}\} && (9.60)\\ &\geq \text{Tr}\{\Pi(\sigma - \rho)\} && (9.61)\\ &= \text{Tr}\{\Pi\sigma\} - \text{Tr}\{\Pi\rho\}. && (9.62) \end{aligned}$$

The first equality follows from Lemma 9.1.1. The first inequality follows from the fact that

$$2 \max_{0 \leq \Lambda \leq I} \{\text{Tr}\{\Lambda(\sigma - \rho)\}\} \geq 0, \tag{9.63}$$

and $2 \geq 1$. The second inequality follows because Λ is the maximizing operator and can only lead to a probability difference greater than that for another operator Π such that $0 \leq \Pi \leq I$. □

The most common way that we employ Corollary 9.1.1 in quantum Shannon theory is in the following scenario. Suppose that a measurement with operator Π succeeds with high probability on a quantum state σ:

$$\text{Tr}\{\Pi\sigma\} \geq 1 - \epsilon, \tag{9.64}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

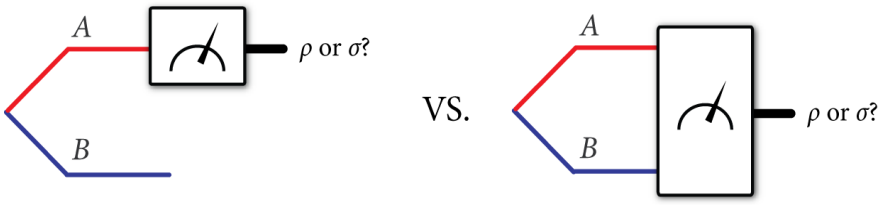

Figure 9.2: The task in the above figure is for Bob to distinguish the state ρ^{AB} from the state σ^{AB} with a binary-valued measurement. Bob could perform an optimal measurement on system A alone if he does not have access to system B. If he has access to system B as well, then he can perform an optimal joint measurement on systems A and B. We would expect that he can distinguish the states more reliably if he performs a joint measurement because there could be more information about the state available in the other system B. Since the trace distance is a measure of distinguishability, we would expect it to obey the following inequality: $\left\|\rho^{A}-\sigma^{A}\right\|_1 \leq \left\|\rho^{AB}-\sigma^{AB}\right\|_1$ (the states are less distinguishable if fewer systems are available to be part of the distinguishability test).

where ϵ is some small positive number. Suppose further that another quantum state ρ is ϵ-close in trace distance to σ:

$$\left\|\rho-\sigma\right\|_1 \leq \epsilon. \tag{9.65}$$

Then Corollary 9.1.1 gives the intuitive result that the measurement succeeds with high probability on the state ρ that is close to σ:

$$\operatorname{Tr}\{\Pi\rho\} \geq 1-2\epsilon, \tag{9.66}$$

by plugging (9.64) and (9.65) into (9.58).

Exercise 9.1.7 Prove that Corollary 9.1.1 holds for arbitrary Hermitian operators ρ and σ by exploiting the result of Exercise 9.1.4.

We next turn to the monotonicity of trace distance under the discarding of a system. The interpretation of this corollary is that discarding of a system does not increase distinguishability of two quantum states. That is, a global measurement on the larger system might be able to distinguish the two states better than a local measurement on an individual subsystem could. In fact, the proof of monotonicity follows this intuition exactly, and Figure 9.2 depicts the intuition behind it.

Corollary 9.1.2 (Monotonicity)**.** *The trace distance is monotone under discarding of subsystems:*

$$\left\|\rho^{A}-\sigma^{A}\right\|_1 \leq \left\|\rho^{AB}-\sigma^{AB}\right\|_1. \tag{9.67}$$

Proof. Consider that

$$\left\|\rho^{A}-\sigma^{A}\right\|_1 = 2\operatorname{Tr}\left\{\Lambda^{A}\left(\rho^{A}-\sigma^{A}\right)\right\}, \tag{9.68}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

for some positive operator Λ^A with spectrum bounded by one. Then

$$\begin{aligned} 2\ \mathrm{Tr}\{\Lambda^A(\rho^A - \sigma^A)\} &= 2\ \mathrm{Tr}\{\Lambda^A \otimes I^B(\rho^{AB} - \sigma^{AB})\} && (9.69)\\ &\leq 2 \max_{0\leq\Lambda^{AB}\leq I} \mathrm{Tr}\{\Lambda^{AB}(\rho^{AB} - \sigma^{AB})\} && (9.70)\\ &= \|\rho^{AB} - \sigma^{AB}\|_1. && (9.71)\end{aligned}$$

The first equality follows because local predictions of the quantum theory should coincide with its global predictions (as discussed in Exercise 4.3.8). The inequality follows because the local operator Λ^A never gives a higher probability difference than a maximization over all global operators. The last equality follows from the characterization of the trace distance in Lemma 9.1.1. □

Exercise 9.1.8 (Monotonicity of Trace Distance under Noisy Maps) Show that the trace distance is monotone under the action of any quantum channel $\mathcal{N}$:

$$\|\mathcal{N}(\rho) - \mathcal{N}(\sigma)\|_1 \leq \|\rho - \sigma\|_1. \tag{9.72}$$

(Hint: Use the result of Corollary 9.1.2 and Exercise 9.1.3.)

The result of the previous exercise deserves an interpretation. It states that a quantum channel $\mathcal{N}$ makes two quantum states ρ and σ less distinguishable from each other. That is, a noisy channel tends to "blur" two states to make them appear as if they are more similar to each other than they are before the quantum channel acts.

Exercise 9.1.9 Show that the trace distance is *strongly convex*. That is, for two ensembles $\{p_{X_1}(x), \rho_x\}$ and $\{p_{X_2}(x), \sigma_x\}$ the following inequality holds

$$\begin{aligned}\left\| \sum_x p_{X_1}(x)\rho_x - \sum_x p_{X_2}(x)\sigma_x \right\|_1 & \\ \leq \sum_x |p_{X_1}(x) - p_{X_2}(x)| + \sum_x p_{X_1}(x)\|\rho_x - \sigma_x\|_1. & \quad (9.73)\end{aligned}$$

9.2 Fidelity

9.2.1 Pure-State Fidelity

An alternate measure of the closeness of two quantum states is the *fidelity*. We introduce its most simple form first. Suppose that we input a particular pure state $|\psi\rangle$ to a quantum information processing protocol. Ideally, we may want the protocol to output the same state that is input, but suppose that it instead outputs a pure state $|\phi\rangle$. The pure-state fidelity $F(|\psi\rangle, |\phi\rangle)$ is a measure of how close the output state is to the input state.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Definition 9.2.1 (Pure-State Fidelity)**.** *The pure-state fidelity is the squared overlap of the states $|\psi\rangle$ and $|\phi\rangle$:*

$$F(|\psi\rangle, |\phi\rangle) \equiv |\langle\psi|\phi\rangle|^2. \tag{9.74}$$

It has the operational interpretation as the probability that the output state $|\phi\rangle$ would pass a test for being the same as the input state $|\psi\rangle$, conducted by someone who knows the input state (see Exercise 9.2.2).

The pure-state fidelity is symmetric

$$F(|\psi\rangle, |\phi\rangle) = F(|\phi\rangle, |\psi\rangle), \tag{9.75}$$

and it obeys the following bounds:

$$0 \leq F(|\psi\rangle, |\phi\rangle) \leq 1. \tag{9.76}$$

It is equal to one if and only if the two states are the same, and it is null if and only if the two states are orthogonal to each other. The fidelity measure is *not* a distance measure in the strict mathematical sense because it is equal to one when two states are equal, whereas a distance measure should be null when two states are equal.

Exercise 9.2.1 Suppose that two quantum states $|\psi\rangle$ and $|\phi\rangle$ are as follows:

$$|\psi\rangle \equiv \sum_x \sqrt{p(x)}|x\rangle, \qquad |\phi\rangle \equiv \sum_x \sqrt{q(x)}|x\rangle, \tag{9.77}$$

where $\{|x\rangle\}$ is some orthonormal basis. Show that the fidelity $F(|\psi\rangle, |\phi\rangle)$ between these two states is equivalent to the *Bhattacharyya distance* between the distributions $p(x)$ and $q(x)$:

$$F(|\psi\rangle, |\phi\rangle) = \left[\sum_x \sqrt{p(x)q(x)}\right]^2. \tag{9.78}$$

9.2.2 Expected Fidelity

Now let us suppose that the output of the protocol is not a pure state, but it is rather a mixed state with density operator ρ. In general, a quantum information processing protocol could be noisy and map the pure input state $|\psi\rangle$ to a mixed state.

Definition 9.2.2 (Expected Fidelity)**.** *The expected fidelity $F(|\psi\rangle, \rho)$ between a pure state $|\psi\rangle$ and a mixed state ρ is*

$$F(|\psi\rangle, \rho) \equiv \langle\psi|\rho|\psi\rangle. \tag{9.79}$$

We now justify the above definition of fidelity. Let us decompose ρ according to its spectral decomposition $\rho = \sum_x p_X(x)|\phi_x\rangle\langle\phi_x|$. Recall that we can think of this output density operator as arising from the ensemble $\{p_X(x), |\phi_x\rangle\}$. We generalize the pure state

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

fidelity from the previous paragraph by defining it as the expected pure state fidelity, where the expectation is with respect to states in the ensemble:

$$\begin{aligned} F(|\psi\rangle, \rho) &\equiv \mathbb{E}_X \left[|\langle \psi | \phi_X \rangle|^2 \right] && (9.80) \\ &= \sum_x p_X(x) |\langle \psi | \phi_x \rangle|^2 && (9.81) \\ &= \sum_x p_X(x) \langle \psi | \phi_x \rangle \langle \phi_x | \psi \rangle && (9.82) \\ &= \langle \psi | \left(\sum_x p_X(x) |\phi_x\rangle\langle\phi_x| \right) |\psi\rangle && (9.83) \\ &= \langle \psi | \rho | \psi \rangle. && (9.84) \end{aligned}$$

The compact formula $F(|\psi\rangle, \rho) = \langle \psi | \rho | \psi \rangle$ is a good way to characterize the fidelity when the input state is pure and the output state is mixed. We can see that the above fidelity measure is a generalization of the pure state fidelity in (9.74). It obeys the same bounds:

$$0 \leq F(\rho, |\psi\rangle) \leq 1, \tag{9.85}$$

being equal to one if and only if the state ρ is equivalent to $|\psi\rangle$ and equal to zero if and only if the support of ρ is orthogonal to $|\phi\rangle$.

Exercise 9.2.2 Given a state σ, we would like to see if it would pass a test for being close to another state $|\varphi\rangle$. We can measure the observable $\{|\varphi\rangle\langle\varphi|, I - |\varphi\rangle\langle\varphi|\}$ with result φ corresponding to a "pass" and the result $I - \varphi$ corresponding to a "fail." Show that the fidelity is then $\Pr\{\text{"pass"}\}$.

Exercise 9.2.3 Using the result of Corollary 9.1.1, show that the following inequality holds for a pure state $|\phi\rangle$ and mixed states ρ and σ:

$$F(\rho, \phi) \leq F(\sigma, \phi) + \|\rho - \sigma\|_1 \tag{9.86}$$

9.2.3 Uhlmann Fidelity

What is the most general form of the fidelity when both quantum states are mixed? We can borrow the above idea of the pure state fidelity that exploits the overlap between two pure states. Suppose that we would like to determine the fidelity between two mixed states ρ^A and σ^A that each live on some quantum system A. Let $|\phi_\rho\rangle^{RA}$ and $|\phi_\sigma\rangle^{RA}$ denote particular respective purifications of the mixed states to some reference system R. We can define the Uhlmann fidelity $F(\rho^A, \sigma^A)$ between two mixed states ρ^A and σ^A as the maximum overlap between their respective purifications, where the maximization is with respect to all purifications $|\phi_\rho\rangle^{RA}$ and $|\phi_\sigma\rangle^{RA}$ of the respective states ρ and σ:

$$F(\rho, \sigma) \equiv \max_{|\phi_\rho\rangle^{RA},\ |\phi_\sigma\rangle^{RA}} \left| \langle \phi_\rho | \phi_\sigma \rangle^{RA} \right|^2. \tag{9.87}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can express the fidelity as a maximization over unitaries instead (recall the result of Exercise 5.1.2 that all purifications are equivalent up to unitaries on the reference system):

$$F(\rho,\sigma) = \max_{U_\rho, U_\sigma} \left| \langle \phi_\rho | \Big(\big(U_\rho^\dagger\big)^R \otimes I^A \Big) \big(U_\sigma^R \otimes I^A \big) | \phi_\sigma \rangle^{RA} \right|^2 \tag{9.88}$$

$$= \max_{U_\rho, U_\sigma} \left| \langle \phi_\rho | \big(U_\rho^\dagger U_\sigma \big)^R \otimes I^A | \phi_\sigma \rangle^{RA} \right|^2. \tag{9.89}$$

It is unnecessary to maximize over two sets of unitaries because the product $U_\rho^\dagger U_\sigma$ represents only a single unitary. The final expression for the fidelity between two mixed states is then defined as the Uhlmann fidelity.

Definition 9.2.3 (Uhlmann Fidelity)**.** *The Uhlmann fidelity $F\big(\rho^A, \sigma^A\big)$ between two mixed states ρ^A and σ^A is the maximum overlap between their respective purifications, where the maximization is with respect to all unitaries U on the purification system R:*

$$F(\rho,\sigma) = \max_U \left| \langle \phi_\rho | U^R \otimes I^A | \phi_\sigma \rangle^{RA} \right|^2. \tag{9.90}$$

We will find that this notion of fidelity generalizes the pure-state fidelity in (9.74) and the expected fidelity in (9.84). This holds because the following formula for the fidelity of two mixed states, characterized in terms of the ℓ_1-norm, is equivalent to the above Uhlmann characterization:

$$F(\rho,\sigma) = \left\| \sqrt{\rho}\sqrt{\sigma} \right\|_1^2. \tag{9.91}$$

We state this result as Uhlmann's theorem.

Theorem 9.2.1 (Uhlmann's Theorem)**.** *The following two expressions for fidelity are equivalent:*

$$F(\rho,\sigma) = \max_U \left| \langle \phi_\rho | U^R \otimes I^A | \phi_\sigma \rangle^{RA} \right|^2 = \left\| \sqrt{\rho}\sqrt{\sigma} \right\|_1^2. \tag{9.92}$$

Proof. We can obtain the state ρ by partial tracing over the system R in the following state:

$$|\phi_\rho\rangle^{RA} \equiv \sqrt{d} \Big(I^R \otimes \sqrt{\rho}^A \Big) |\Phi\rangle^{RA}, \tag{9.93}$$

where $\sqrt{\rho}^A$ is an operator acting on the A system and $|\Phi\rangle^{RA}$ is the maximally entangled state with respect to some basis $\{|i\rangle\}$:

$$|\Phi\rangle^{RA} \equiv \frac{1}{\sqrt{d}} \sum_{x=1}^{d} |i\rangle^R |i\rangle^A. \tag{9.94}$$

Therefore, the state $|\phi_\rho\rangle^{RA}$ is a particular purification of ρ. Consider another state σ^A. We can also obtain σ^A as a partial trace over the R system of the following state:

$$|\phi_\sigma\rangle^{RA} \equiv \sqrt{d} \Big(I^R \otimes \sqrt{\sigma}^A \Big) |\Phi\rangle^{RA}, \tag{9.95}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and so $|\phi_\sigma\rangle$ is a purification of σ. Consider that the overlap $\left|\langle\phi_\rho|U^R \otimes I^A|\phi_\sigma\rangle\right|^2$ is as follows:

$$|\langle\phi_\rho|\phi_\sigma\rangle|^2 = \left|d\langle\Phi|^{RA}\Big(U^R \otimes \sqrt{\rho}^A\Big)\Big(I^R \otimes \sqrt{\sigma}^A\Big)|\Phi\rangle^{RA}\right|^2 \tag{9.96}$$

$$= \left|\sum_{i,j}\langle i|^R\langle i|^A\Big(U^R \otimes \big(\sqrt{\rho}\sqrt{\sigma}\big)^A\Big)|j\rangle^R|j\rangle^A\right|^2 \tag{9.97}$$

$$= \left|\sum_{i,j}\langle i|^R\langle i|^A\Big(I^R \otimes \big(\sqrt{\rho}\sqrt{\sigma}U^T\big)^A\Big)|j\rangle^R|j\rangle^A\right|^2 \tag{9.98}$$

$$= \left|\sum_{i,j}\langle i|j\rangle^R\langle i|^A\big(\sqrt{\rho}\sqrt{\sigma}U^T\big)^A|j\rangle^A\right|^2 \tag{9.99}$$

$$= \left|\sum_{i}\langle i|\sqrt{\rho}\sqrt{\sigma}U^T|i\rangle\right|^2 \tag{9.100}$$

$$= \left|\mathrm{Tr}\big\{\sqrt{\rho}\sqrt{\sigma}U^T\big\}\right|^2. \tag{9.101}$$

The first equality follows by plugging (9.93) and (9.95) into the overlap expression $|\langle\phi_\rho|\phi_\sigma\rangle|^2$. The second through fifth equalities follow by evaluation. The last equality follows by the definition of the trace. Using the result of Theorem A.0.3 in Appendix A, it holds that

$$|\langle\phi_\rho|\phi_\sigma\rangle|^2 = \left|\mathrm{Tr}\big\{\sqrt{\rho}\sqrt{\sigma}U^T\big\}\right|^2 \tag{9.102}$$

$$\leq \mathrm{Tr}\big\{\big|\sqrt{\rho}\sqrt{\sigma}\big|\big\}^2 \tag{9.103}$$

$$= \big\|\sqrt{\rho}\sqrt{\sigma}\big\|_1^2. \tag{9.104}$$

Choosing U^T as the inverse of the unitary in the right polar decomposition of $\sqrt{\rho}\sqrt{\sigma}$ saturates the upper bound above. This unitary U is also the maximal unitary in the Uhlmann fidelity in (9.90). □

9.2.4 Properties of Fidelity

We discuss some further properties of the fidelity that often prove useful. Some of these properties are the counterpart of similar properties of the trace distance. From the characterization of fidelity in (9.92), we observe that it is symmetric in its arguments:

$$F(\rho,\sigma) = F(\sigma,\rho). \tag{9.105}$$

It obeys the following bounds:

$$0 \leq F(\rho,\sigma) \leq 1. \tag{9.106}$$

The lower bound applies if and only if the respective supports of the two states ρ and σ are orthogonal. The upper bound applies if and only if the two states ρ and σ are equal to each other.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Property 9.2.1 (Multiplicativity over tensor products) The fidelity is multiplicative over tensor products:

$$F(\rho_1 \otimes \rho_2, \sigma_1 \otimes \sigma_2) = F(\rho_1, \sigma_1) F(\rho_2, \sigma_2). \tag{9.107}$$

This result holds by employing the definition of the fidelity in (9.91).

Property 9.2.2 (Joint Concavity) The fidelity is jointly concave in its input arguments:

$$F\left(\sum_x p_X(x)\rho_x, \sum_x p_X(x)\sigma_x\right) \geq \sum_x p_X(x) F(\rho_x, \sigma_x). \tag{9.108}$$

Proof. We prove joint concavity by exploiting the result of Exercise 5.1.4. Suppose $|\phi_{\rho_x}\rangle^{RA}$ and $|\phi_{\sigma_x}\rangle^{RA}$ are respective Uhlmann purifications of ρ_x and σ_x (these are purifications that maximize the Uhlmann fidelity). Then

$$F\big(|\phi_{\rho_x}\rangle^{RA}, |\phi_{\sigma_x}\rangle^{RA}\big) = F(\rho_x, \sigma_x). \tag{9.109}$$

Choose some orthonormal basis $\left\{|x\rangle^X\right\}$. Then

$$|\phi_\rho\rangle \equiv \sum_x \sqrt{p_X(x)} |\phi_{\rho_x}\rangle^{RA} |x\rangle^X, \qquad |\phi_\sigma\rangle \equiv \sum_x \sqrt{p_X(x)} |\phi_{\sigma_x}\rangle^{RA} |x\rangle^X \tag{9.110}$$

are respective purifications of $\sum_x p_X(x)\rho_x$ and $\sum_x p_X(x)\sigma_x$. The first inequality below holds by Uhlmann's theorem:

$$\begin{aligned} F\left(\sum_x p_X(x)\rho_x, \sum_x p_X(x)\sigma_x\right) &\geq |\langle \phi_\rho | \phi_\sigma \rangle|^2 && (9.111)\\ &= \sum_x p_X(x) |\langle \phi_{\rho_x} | \phi_{\sigma_x} \rangle|^2 && (9.112)\\ &= \sum_x p_X(x) F(\rho_x, \sigma_x). && (9.113) \end{aligned}$$

□

Property 9.2.3 (Concavity) The fidelity is concave over one of its arguments

$$F(\lambda\rho_1 + (1-\lambda)\rho_2, \sigma) \geq \lambda F(\rho_1, \sigma) + (1-\lambda) F(\rho_2, \sigma). \tag{9.114}$$

Concavity follows from joint concavity (Property 9.2.2).

The following monotonicity lemma is similar to the monotonicity lemma for trace distance (Lemma 9.1.2) and also bears the similar interpretation that quantum states become more similar (less distinguishable) under the discarding of subsystems.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Lemma 9.2.1 (Monotonicity). *The fidelity is non-decreasing under partial trace:*

$$F\big(\rho^{AB},\sigma^{AB}\big) \leq F\big(\rho^{A},\sigma^{A}\big), \tag{9.115}$$

where

$$\rho^{A} = \text{Tr}_{B}\big\{\rho^{AB}\big\}, \qquad \sigma^{A} = \text{Tr}_{B}\big\{\sigma^{AB}\big\}. \tag{9.116}$$

Proof. Consider a fixed purification $|\psi\rangle^{RAB}$ of ρ^{A} and ρ^{AB} and a fixed purification $|\phi\rangle^{RAB}$ of σ^{A} and σ^{AB}. By Uhlmann's theorem,

$$F\big(\rho^{AB},\sigma^{AB}\big) = \max_{U^{R}\otimes I^{AB}} \Big|\big\langle\psi|U^{R}\otimes I^{AB}|\phi\big\rangle^{RAB}\Big|^{2}, \tag{9.117}$$

$$F\big(\rho^{A},\sigma^{A}\big) = \max_{U^{RB}\otimes I^{A}} \Big|\big\langle\psi|U^{RB}\otimes I^{A}|\phi\big\rangle^{RAB}\Big|^{2}. \tag{9.118}$$

The inequality in the statement of the theorem then holds because the maximization of $\Big|\big\langle\psi|U^{RB}\otimes I^{A}|\phi\big\rangle^{RAB}\Big|^{2}$ for $F\big(\rho^{A},\sigma^{A}\big)$ is inclusive of all the unitaries in the maximization of $\Big|\big\langle\psi|U^{R}\otimes I^{AB}|\phi\big\rangle^{RAB}\Big|^{2}$ for $F\big(\rho^{AB},\sigma^{AB}\big)$. Thus, $F\big(\rho^{A},\sigma^{A}\big)$ can only be larger or equal to $F\big(\rho^{AB},\sigma^{AB}\big)$. □

Exercise 9.2.4 Show that we can express the fidelity as

$$F(\rho,\sigma) = \text{Tr}\left\{\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right\}^{2}, \tag{9.119}$$

using the definition in (9.91).

Exercise 9.2.5 Show that the fidelity is invariant under an isometry U:

$$F(\rho,\sigma) = F\big(U\rho U^{\dagger}, U\sigma U^{\dagger}\big). \tag{9.120}$$

Exercise 9.2.6 Show that the fidelity is monotone under a noisy quantum operation $\mathcal{N}$:

$$F(\rho,\sigma) \leq F(\mathcal{N}(\rho),\mathcal{N}(\sigma)). \tag{9.121}$$

Exercise 9.2.7 Suppose that Alice uses a noisy quantum channel and a sequence of quantum operations to generate the following state, shared with Bob and Eve:

$$\frac{1}{\sqrt{M}}\sum_{m}|m\rangle^{A}|m\rangle^{B_1}|\phi_m\rangle^{B_2E}, \tag{9.122}$$

where Alice possesses the system A, Bob possesses systems B_1 and B_2, and Eve possesses the system E. Let ϕ_m^E denote the partial trace of $|\phi_m\rangle^{B_2E}$ over Bob's system B_2 so that

$$\phi_m^E \equiv \text{Tr}_{B_2}\Big\{|\phi_m\rangle\langle\phi_m|^{B_2E}\Big\}. \tag{9.123}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

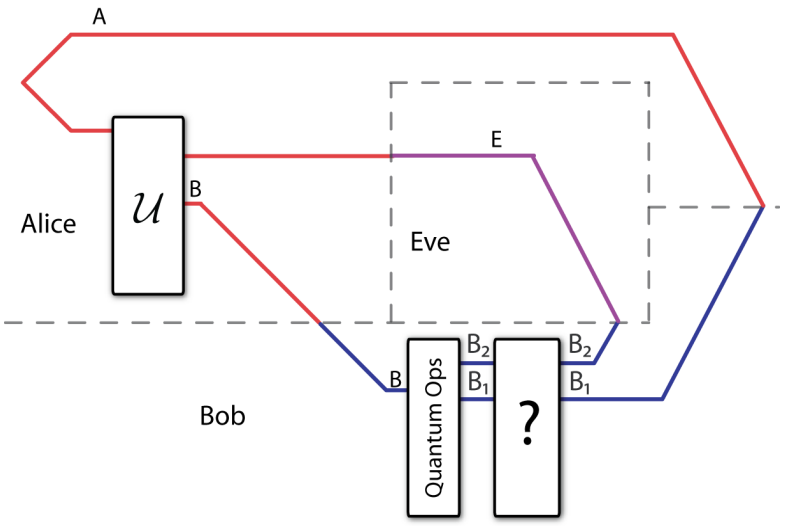

Figure 9.3: The above figure depicts the protocol relevant to Exercise 9.2.7. Alice transmits one share of an entangled state through a noisy quantum channel with isometric extension $\mathcal{U}$. Bob and Eve receive quantum systems as the output of the isometry. Bob performs some quantum operations so that Alice, Bob, and Eve share the state in (9.122). Exercise 9.2.7 asks you to determine a decoupling unitary that Bob can perform to decouple his system B_1 from Eve.

Suppose further that $F(\phi_m^E, \theta^E) = 1$, where θ^E is some *constant* density operator (independent of m) living on Eve's system E. Determine a unitary that Bob can perform on his systems B_1 and B_2 so that he *decouples* Eve's system E, in the sense that the state after the decoupling unitary is as follows:

$$\left(\frac{1}{\sqrt{M}} \sum_m |m\rangle^A |m\rangle^{B_1} \right) \otimes |\phi_\theta\rangle^{B_2 E}, \tag{9.124}$$

where $|\phi_\theta\rangle^{B_2 E}$ is a purification of the state θ^E. The result is that Alice and Bob share maximal entanglement between the respective systems A and B_1 after Bob performs the decoupling unitary. Figure 9.3 displays the protocol.

9.3 Relationships between Trace Distance and Fidelity

In quantum Shannon theory, we are interested in showing that a given quantum information processing protocol approximates an ideal protocol. We might do so by showing that the quantum output of the ideal protocol, say ρ, is approximately close to the quantum output of the actual protocol, say σ. For example, we may be able to show that the fidelity between ρ and σ is high:

$$F(\rho, \sigma) \geq 1 - \epsilon, \tag{9.125}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where ϵ is a small, positive real number that determines how well ρ approximates σ according to the above fidelity criterion. Typically, in a quantum Shannon-theoretic argument, we will take a limit to show that it is possible to make ϵ as small as we would like. As the performance paramater ϵ becomes vanishingly small, we expect that ρ and σ are becoming approximately equal so that they are identically equal when ϵ vanishes in some limit.

We would naturally think that the trace distance should be small if the fidelity is high because the trace distance vanishes when the fidelity is one and vice versa (recall the conditions for saturation of the bounds in (9.11) and (9.106)). The next theorem makes this intuition precise by establishing several relationships between the trace distance and fidelity.

Theorem 9.3.1 (Relations between Fidelity and Trace Distance)**.** *The following bound applies to the trace distance and the fidelity between two quantum states ρ and σ:*

$$1 - \sqrt{F(\rho,\sigma)} \leq \frac{1}{2}\|\rho - \sigma\|_1 \leq \sqrt{1 - F(\rho,\sigma)}. \tag{9.126}$$

Proof. We first show that there is an exact relationship between fidelity and trace distance for pure states. Let us pick two arbitrary pure states $|\psi\rangle$ and $|\phi\rangle$. We can write the state $|\phi\rangle$ in terms of the state $|\psi\rangle$ and its orthogonal complement $|\psi^\perp\rangle$ as follows:

$$|\phi\rangle = \cos(\theta)|\psi\rangle + e^{i\varphi}\sin(\theta)|\psi^\perp\rangle. \tag{9.127}$$

First, the fidelity between these two pure states is

$$F(|\psi\rangle, |\phi\rangle) = |\langle\phi|\psi\rangle|^2 = \cos^2(\theta). \tag{9.128}$$

Now let us determine the trace distance. The density operator $|\phi\rangle\langle\phi|$ is as follows:

$$\begin{aligned} |\phi\rangle\langle\phi| &= \left(\cos(\theta)|\psi\rangle + e^{i\varphi}\sin(\theta)|\psi^\perp\rangle\right)\left(\cos(\theta)\langle\psi| + e^{-i\varphi}\sin(\theta)\langle\psi^\perp|\right) &(9.129)\\ &= \cos^2(\theta)|\psi\rangle\langle\psi| + e^{i\varphi}\sin(\theta)\cos(\theta)|\psi^\perp\rangle\langle\psi| \\ &\quad + e^{-i\varphi}\cos(\theta)\sin(\theta)|\psi\rangle\langle\psi^\perp| + \sin^2(\theta)|\psi^\perp\rangle\langle\psi^\perp|. &(9.130)\end{aligned}$$

The matrix representation of the operator $|\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|$ with respect to the basis $\{|\psi\rangle, |\psi^\perp\rangle\}$ is

$$\begin{bmatrix} 1-\cos^2(\theta) & -e^{-i\varphi}\sin(\theta)\cos(\theta) \\ -e^{i\varphi}\sin(\theta)\cos(\theta) & -\sin^2(\theta) \end{bmatrix}. \tag{9.131}$$

It is straightforward to show that the eigenvalues of the above matrix are $|\sin(\theta)|$ and $-|\sin(\theta)|$ and it then follows that the trace distance between $|\psi\rangle$ and $|\phi\rangle$ is the absolute sum of the eigenvalues:

$$\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_1 = 2|\sin(\theta)|. \tag{9.132}$$

Consider the following trigonometric relationship:

$$\left(\frac{2|\sin(\theta)|}{2}\right)^2 = 1 - \cos^2(\theta). \tag{9.133}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

It then holds that the fidelity and trace distance for pure states are related as follows:

$$\left(\frac{1}{2}\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_1\right)^2 = 1 - F(|\psi\rangle, |\phi\rangle), \tag{9.134}$$

by plugging (9.128) into the RHS of (9.133) and (9.132) into the LHS of (9.133). Thus,

$$\frac{1}{2}\||\psi\rangle\langle\psi| - |\phi\rangle\langle\phi|\|_1 = \sqrt{1 - F(|\psi\rangle, |\phi\rangle)}. \tag{9.135}$$

To prove the upper bound for mixed states ρ^A and σ^A, choose purifications $|\phi_\rho\rangle^{RA}$ and $|\phi_\sigma\rangle^{RA}$ of respective states ρ^A and σ^A such that

$$F\left(\rho^A, \sigma^A\right) = |\langle\phi_\sigma|\phi_\rho\rangle|^2 = F\left(|\phi_\rho\rangle^{RA}, |\phi_\sigma\rangle^{RA}\right). \tag{9.136}$$

(Recall that these purifications exist by Uhlmann's theorem.) Then

$$\frac{1}{2}\left\|\rho^A - \sigma^A\right\|_1 \leq \frac{1}{2}\left\|\phi_\rho^{RA} - \phi_\sigma^{RA}\right\|_1 \tag{9.137}$$

$$= \sqrt{1 - F\left(|\phi_\rho\rangle^{RA}, |\phi_\sigma\rangle^{RA}\right)} \tag{9.138}$$

$$= \sqrt{1 - F(\rho^A, \sigma^A)}, \tag{9.139}$$

where the first inequality follows by the monotonicity of the trace distance under the discarding of systems (Lemma 9.1.2). To prove the lower bound for mixed states ρ and σ, we first state the following theorems without proof. It is possible to show that the trace distance is the maximum Kolmogorov distance between two probability distributions resulting from a POVM $\{\Lambda_m\}$ acting on the states ρ and σ:

$$\|\rho - \sigma\|_1 = \max_{\{\Lambda_m\}} \sum_m |p_m - q_m|, \tag{9.140}$$

where

$$p_m \equiv \mathrm{Tr}\{\Lambda_m \rho\}, \qquad q_m \equiv \mathrm{Tr}\{\Lambda_m \sigma\}. \tag{9.141}$$

It is also possible to show that the fidelity is the minimum Bhattacharya distance between two probability distributions p'_m and q'_m resulting from a measurement $\{\Gamma_m\}$ of the states ρ and σ:

$$F(\rho, \sigma) = \min_{\{\Gamma_m\}} \left(\sum_m \sqrt{p'_m q'_m}\right)^2, \tag{9.142}$$

where

$$p'_m \equiv \mathrm{Tr}\{\Gamma_m \rho\}, \qquad q'_m \equiv \mathrm{Tr}\{\Gamma_m \sigma\}. \tag{9.143}$$

We return to the proof. Suppose that the POVM Γ_m achieves the minimum Bhattacharya distance and results in probability distributions p'_m and q'_m, so that

$$F(\rho, \sigma) = \left(\sum_m \sqrt{p'_m q'_m}\right)^2. \tag{9.144}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Consider that

$$\sum_m \left(\sqrt{p'_m} - \sqrt{q'_m}\right)^2 = \sum_m p'_m + q'_m - \sqrt{p'_m q'_m} \tag{9.145}$$

$$= 2 - 2\sqrt{F(\rho,\sigma)} \tag{9.146}$$

It also follows that

$$\sum_m \left(\sqrt{p'_m} - \sqrt{q'_m}\right)^2 \leq \sum_m \left|\sqrt{p'_m} - \sqrt{q'_m}\right| \left|\sqrt{p'_m} + \sqrt{q'_m}\right| \tag{9.147}$$

$$= \sum_m |p'_m - q'_m| \tag{9.148}$$

$$\leq \sum_m |p_m - q_m| \tag{9.149}$$

$$= \|\rho - \sigma\|_1. \tag{9.150}$$

The first inequality holds because $|\sqrt{p'_m} - \sqrt{q'_m}| \leq |\sqrt{p'_m} + \sqrt{q'_m}|$. The second inequality holds because the distributions p'_m and q'_m minimizing the Bhattacharya distance in general have Kolmogorov distance less than the distributions p_m and q_m that maximize the Kolmogorov distance. Thus, the following inequality results

$$2 - 2\sqrt{F(\rho,\sigma)} \leq \|\rho - \sigma\|_1, \tag{9.151}$$

and the lower bound in the statement of the theorem follows. □

The following two corollaries are simple consequences of Theorem 9.3.1.

Corollary 9.3.1. *Suppose that ρ is ϵ-close to σ in trace distance:*

$$\|\rho - \sigma\|_1 \leq \epsilon. \tag{9.152}$$

Then the fidelity between ρ and σ is greater than $1 - \epsilon$:

$$F(\rho,\sigma) \geq 1 - \epsilon. \tag{9.153}$$

Corollary 9.3.2. *Suppose the fidelity between ρ and σ is greater than $1 - \epsilon$:*

$$F(\rho,\sigma) \geq 1 - \epsilon. \tag{9.154}$$

Then ρ is $2\sqrt{\epsilon}$-close to σ in trace distance:

$$\|\rho - \sigma\|_1 \leq 2\sqrt{\epsilon}. \tag{9.155}$$

Exercise 9.3.1 Prove the following lower bound on the probability of error P_e in a quantum hypothesis test to distinguish ρ from σ:

$$P_e \geq \frac{1}{2}\left(1 - \sqrt{1 - F(\rho,\sigma)}\right). \tag{9.156}$$

(Hint: Recall the development in Section 9.1.4.)

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

9.4 Gentle Measurement

The Gentle Measurement and Gentle Operator Lemmas are particular applications of Theorem 9.3.1, and they concern the disturbance of quantum states. We generally expect in quantum theory that certain measurements might disturb the state which we are measuring. For example, suppose a qubit is in the state $|0\rangle$. A measurement along the X axis gives $+1$ and -1 with equal probability while drastically disturbing the state to become either $|+\rangle$ or $|-\rangle$, respectively. On the other hand, we might expect that the measurement does not disturb the state by very much if one outcome is highly likely. For example, suppose that we instead measure the qubit along the Z axis. The measurement returns $+1$ with unit probability while causing no disturbance to the qubit. The below "Gentle Measurement Lemma" quantitatively addresses the disturbance of quantum states by demonstrating that a measurement with one outcome that is highly likely causes only a little disturbance to the quantum state that we measure (hence, the measurement is "gentle" or "tender").

Lemma 9.4.1 (Gentle Measurement)**.** *Consider a density operator ρ and a measurement operator Λ where $0 \leq \Lambda \leq I$. The measurement operator could be an element of a POVM. Suppose that the measurement operator Λ has a high probability of detecting state ρ:*

$$\text{Tr}\{\Lambda\rho\} \geq 1 - \epsilon, \tag{9.157}$$

where $1 \geq \epsilon > 0$ (the probability of detection is high only if ϵ is close to zero). Then the post-measurement state

$$\rho' \equiv \frac{\sqrt{\Lambda}\rho\sqrt{\Lambda}}{\text{Tr}\{\Lambda\rho\}} \tag{9.158}$$

is $2\sqrt{\epsilon}$-close to the original state ρ in trace distance:

$$\|\rho - \rho'\|_1 \leq 2\sqrt{\epsilon}. \tag{9.159}$$

Thus, the measurement does not disturb the state ρ by much if ϵ is small.

Proof. Suppose first that ρ is a pure state $|\psi\rangle\langle\psi|$. The post-measurement state is then

$$\frac{\sqrt{\Lambda}|\psi\rangle\langle\psi|\sqrt{\Lambda}}{\langle\psi|\Lambda|\psi\rangle}. \tag{9.160}$$

The fidelity between the original state $|\psi\rangle$ and the post-measurement state above is as follows:

$$\langle\psi|\left(\frac{\sqrt{\Lambda}|\psi\rangle\langle\psi|\sqrt{\Lambda}}{\langle\psi|\Lambda|\psi\rangle}\right)|\psi\rangle = \frac{\left|\langle\psi|\sqrt{\Lambda}|\psi\rangle\right|^2}{\langle\psi|\Lambda|\psi\rangle} \tag{9.161}$$

$$\geq \frac{|\langle\psi|\Lambda|\psi\rangle|^2}{\langle\psi|\Lambda|\psi\rangle} \tag{9.162}$$

$$= \langle\psi|\Lambda|\psi\rangle \tag{9.163}$$

$$\geq 1 - \epsilon. \tag{9.164}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first inequality follows because $\sqrt{\Lambda} \geq \Lambda$ when $\Lambda \leq I$. The second inequality follows from the hypothesis of the lemma. Now let us consider when we have mixed states ρ^A and ρ'^A. Suppose $|\psi\rangle^{RA}$ and $|\psi'\rangle^{RA}$ are respective purifications of ρ^A and ρ'^A, where

$$|\psi'\rangle^{RA} \equiv \frac{I^R \otimes \sqrt{\Lambda}^A |\psi\rangle^{RA}}{\sqrt{\langle\psi| I^R \otimes \Lambda^A |\psi\rangle^{RA}}}. \tag{9.165}$$

Then we can apply monotonicity of fidelity (Lemma 9.2.1) and the above result for pure states to show that

$$F\big(\rho^A, \rho'^A\big) \geq F\Big(|\psi\rangle^{RA}, |\psi'\rangle^{RA}\Big) \geq 1 - \epsilon. \tag{9.166}$$

We finally obtain the bound on the trace distance $\big\|\rho^A - \rho'^A\big\|_1$ by exploiting Corollary 9.3.2. □

The following lemma is a variation on the Gentle Measurement Lemma that we sometimes exploit.

Lemma 9.4.2 (Gentle Operator)**.** *Consider a density operator ρ and a measurement operator Λ where $0 \leq \Lambda \leq I$. The measurement operator could be an element of a POVM. Suppose that the measurement operator Λ has a high probability of detecting state ρ:*

$$\mathrm{Tr}\{\Lambda\rho\} \geq 1 - \epsilon, \tag{9.167}$$

where $1 \geq \epsilon > 0$ (the probability is high only if ϵ is close to zero). Then $\sqrt{\Lambda}\rho\sqrt{\Lambda}$ is $2\sqrt{\epsilon}$-close to the original state ρ in trace distance:

$$\Big\|\rho - \sqrt{\Lambda}\rho\sqrt{\Lambda}\Big\|_1 \leq 2\sqrt{\epsilon}. \tag{9.168}$$

Proof. Consider the following chain of inequalities:

$$\begin{aligned}
&\Big\|\rho - \sqrt{\Lambda}\rho\sqrt{\Lambda}\Big\|_1 \\
&= \Big\|\Big(I - \sqrt{\Lambda} + \sqrt{\Lambda}\Big)\rho - \sqrt{\Lambda}\rho\sqrt{\Lambda}\Big\|_1 && (9.169)\\
&\leq \Big\|\Big(I - \sqrt{\Lambda}\Big)\rho\Big\|_1 + \Big\|\sqrt{\Lambda}\rho\Big(I - \sqrt{\Lambda}\Big)\Big\|_1 && (9.170)\\
&= \mathrm{Tr}\Big|\Big(I - \sqrt{\Lambda}\Big)\sqrt{\rho}\cdot\sqrt{\rho}\Big| + \mathrm{Tr}\Big|\sqrt{\Lambda}\sqrt{\rho}\cdot\sqrt{\rho}\Big(I - \sqrt{\Lambda}\Big)\Big| && (9.171)\\
&\leq \sqrt{\mathrm{Tr}\Big\{\Big(I - \sqrt{\Lambda}\Big)^2\rho\Big\}\mathrm{Tr}\{\rho\}} + \sqrt{\mathrm{Tr}\{\Lambda\rho\}\mathrm{Tr}\Big\{\rho\Big(I - \sqrt{\Lambda}\Big)^2\Big\}} && (9.172)\\
&\leq \sqrt{\mathrm{Tr}\{(I - \Lambda)\rho\}} + \sqrt{\mathrm{Tr}\{\rho(I - \Lambda)\}} && (9.173)\\
&= 2\sqrt{\mathrm{Tr}\{(I - \Lambda)\rho\}} && (9.174)\\
&\leq 2\sqrt{\epsilon}. && (9.175)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first inequality is the triangle inequality. The second equality follows from the definition of the trace norm and the fact that ρ is a positive operator. The second inequality is the Cauchy-Schwarz inequality for the Hilbert-Schmidt norm and any two operators A and B:

$$\text{Tr}\{A^\dagger B\} \leq \sqrt{\text{Tr}\{A^\dagger A\}\text{Tr}\{B^\dagger B\}}. \tag{9.176}$$

The third inequality follows because $(1-\sqrt{x})^2 \leq 1-x$ for $0 \leq x \leq 1$, $\text{Tr}\{\rho\} = 1$, and $\text{Tr}\{\Lambda\rho\} \leq 1$. The final inequality follows from applying (9.167) and because the square root function is monotone increasing. □

Exercise 9.4.1 Show that the Gentle Operator Lemma holds for subnormalized positive operators ρ (operators ρ such that $\text{Tr}\{\rho\} \leq 1$).

Below is another variation on the Gentle Measurement Lemma that applies to ensembles of quantum states.

Lemma 9.4.3 (Gentle Measurement for Ensembles)**.** *Let $\{p_x, \rho_x\}$ be an ensemble with average $\overline{\rho} \equiv \sum_x p_x \rho_x$. Given a positive operator Λ with $\Lambda \leq I$ and $Tr\{\overline{\rho}\Lambda\} \geq 1-\epsilon$ where $\epsilon \leq 1$, then*

$$\sum_x p_x \left\| \rho_x - \sqrt{\Lambda}\rho_x\sqrt{\Lambda} \right\|_1 \leq 2\sqrt{\epsilon}. \tag{9.177}$$

Proof. We can apply the same steps in the proof of the Gentle Operator Lemma to get the following inequality:

$$\left\| \rho_x - \sqrt{\Lambda}\rho_x\sqrt{\Lambda} \right\|_1^2 \leq 4(1 - \text{Tr}\{\rho_x\Lambda\}). \tag{9.178}$$

Taking the expectation over both sides produces the following inequality:

$$\sum_x p_x \left\| \rho_x - \sqrt{\Lambda}\rho_x\sqrt{\Lambda} \right\|_1^2 \leq 4(1 - \text{Tr}\{\rho\Lambda\}) \tag{9.179}$$

$$\leq 4\epsilon. \tag{9.180}$$

Taking the square root of the above inequality gives the following one:

$$\sqrt{\sum_x p_x \left\| \rho_x - \sqrt{\Lambda}\rho_x\sqrt{\Lambda} \right\|_1^2} \leq 2\sqrt{\epsilon}. \tag{9.181}$$

Concavity of the square root implies then implies the result:

$$\sum_x p_x \sqrt{\left\| \rho_x - \sqrt{\Lambda}\rho_x\sqrt{\Lambda} \right\|_1^2} \leq 2\sqrt{\epsilon}. \tag{9.182}$$

□

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 9.4.2 (Coherent Gentle Measurement) Let $\{\rho_k^A\}$ be a collection of density operators and $\{\Lambda_k\}$ be a POVM such that for all k:

$$\text{Tr}\{\Lambda_k^A \rho_k^A\} \geq 1 - \epsilon. \tag{9.183}$$

Let $|\phi_k\rangle^{RA}$ be a purification of ρ_k^A. Show that there exists a coherent gentle measurement $\mathcal{D}^{A\to AK}$ in the sense of Section 5.4 such that

$$\left\| \mathcal{D}^{A\to AK}(\phi_k^{RA}) - \phi_k^{RA} \otimes |k\rangle\langle k|^K \right\| \leq 2\sqrt{\epsilon}. \tag{9.184}$$

(Hint: Use the result of Exercise 5.4.1.)

9.5 Fidelity of a Noisy Quantum Channel

It is useful to have measures that determine how well a noisy quantum channel $\mathcal{N}$ preserves quantum information. We developed static distance measures, such as the trace distance and the fidelity, in the previous sections of this chapter. We would now like to exploit those measures in order to define dynamic measures.

A "first guess" measure of this sort is the minimum fidelity $F_{\min}(\mathcal{N})$ where

$$F_{\min}(\mathcal{N}) \equiv \min_{|\psi\rangle} F(|\psi\rangle, \mathcal{N}(|\psi\rangle\langle\psi|)). \tag{9.185}$$

This measure seems like it may be a good one because we generally do not know the state that Alice inputs to a noisy channel before transmitting to Bob.

It may seem somewhat strange that we chose to minimize over pure states in the definition of the minimum fidelity. Are not mixed states the most general states that occur in the quantum theory? It turns out that joint concavity of the fidelity (Property 9.2.2) implies that we do not have to consider mixed states for the minimum fidelity. Consider the following sequence of inequalities:

$$F(\rho, \mathcal{N}(\rho)) = F\left(\sum_x p_X(x)|x\rangle\langle x|, \mathcal{N}\left(\sum_x p_X(x)|x\rangle\langle x|\right)\right) \tag{9.186}$$

$$= F\left(\sum_x p_X(x)|x\rangle\langle x|, \sum_x p_X(x)\mathcal{N}(|x\rangle\langle x|)\right) \tag{9.187}$$

$$\geq \sum_x p_X(x) F(|x\rangle\langle x|, \mathcal{N}(|x\rangle\langle x|)) \tag{9.188}$$

$$\geq F(|x_{\min}\rangle\langle x_{\min}|, \mathcal{N}(|x_{\min}\rangle\langle x_{\min}|)). \tag{9.189}$$

The first equality follows by expanding the density operator ρ with the spectral decomposition. The second equality follows from linearity of the quantum operation $\mathcal{N}$. The first inequality follows from joint concavity of the fidelity (Property 9.2.2), and the last inequality follows because there exists some pure state $|x_{\min}\rangle$ (one of the eigenstates of ρ) with fidelity never larger than the expected fidelity in the previous line.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

9.5.1 Expected Fidelity

In general, the minimum fidelity is less useful than other measures of quantum information preservation over a noisy channel. The difficulty with the minimum fidelity is that it requires an optimization over the potentially large space of input states. Since it is somewhat difficult to manipulate and compute in general, we introduce other ways to determine the performance of a noisy quantum channel.

We can simplify our notion of fidelity by instead restricting the states that Alice sends and averaging the fidelity over this set of states. That is, suppose that Alice is transmitting states from an ensemble $\{p_X(x), \rho_x\}$ and we would like to determine how well a noisy quantum channel $\mathcal{N}$ is preserving this source of quantum information. Sending a particular state ρ_x through a noisy quantum channel $\mathcal{N}$ produces the state $\mathcal{N}(\rho_x)$. The fidelity between the transmitted state ρ_x and the received state $\mathcal{N}(\rho_x)$ is $F(\rho_x, \mathcal{N}(\rho_x))$ as defined before. We define the *expected fidelity* of the ensemble as follows:

$$\overline{F}(\mathcal{N}) \equiv \mathbb{E}_X[F(\rho_X, \mathcal{N}(\rho_X))] \tag{9.190}$$

$$= \sum_x p_X(x) F(\rho_x, \mathcal{N}(\rho_x)). \tag{9.191}$$

The expected fidelity indicates how well Alice is able to transmit the ensemble on average to Bob. It again lies between zero and one, just as the usual fidelity does.

A more general form of the expected fidelity is to consider the expected performance for any quantum state instead of restricting ourselves to an ensemble. That is, let us fix some quantum state $|\psi\rangle$ and apply a random unitary U to it, where we select the unitary according to the Haar measure (this is the uniform distribution on unitaries). The state $U|\psi\rangle$ represents a random quantum state and we can take the expectation over it in order to define the following more general notion of expected fidelity:

$$\overline{F}(\mathcal{N}) \equiv \mathbb{E}_U\left[F\big(U|\psi\rangle, \mathcal{N}\big(U|\psi\rangle\langle\psi|U^\dagger\big)\big)\right], \tag{9.192}$$

The above formula for the expected fidelity then becomes the following integral over the Haar measure:

$$\overline{F}(\mathcal{N}) = \int \langle\psi|U^\dagger \mathcal{N}\big(U|\psi\rangle\langle\psi|U^\dagger\big)U|\psi\rangle \; dU. \tag{9.193}$$

9.5.2 Entanglement Fidelity

We now consider a different measure of the ability of a noisy quantum channel to preserve quantum information. Suppose that Alice would like to transmit a quantum state with density operator ρ^A. It admits a purification $|\psi\rangle^{RA}$ to a reference system R. Sending the A system of $|\psi\rangle^{RA}$ through the identity channel $I^{A\to B}$ gives a state $|\psi\rangle^{RB}$ where B is a system that is isomorphic to A. Sending the A system of $|\psi\rangle^{RA}$ through quantum channel $\mathcal{N}$ gives the state $\sigma^{RB} \equiv \big(I^R \otimes \mathcal{N}^{A\to B}\big)\big(\psi^{RA}\big)$. The entanglement fidelity is as follows:

$$F_e(\rho, \mathcal{N}) \equiv \langle\psi|\sigma|\psi\rangle, \tag{9.194}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

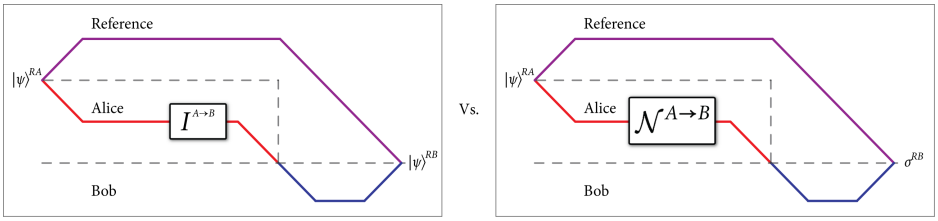

Figure 9.4: The entanglement fidelity compares the output of the ideal scenario (depicted on the left) and the output of the noisy scenario (depicted on the right).

and it is a measure of how well the noisy channel can preserve the entanglement with another system. Figure 9.4 visually depicts the two states that the entanglement fidelity compares.

One of the benefits of considering the task of entanglement preservation is that it implies the task of quantum communication. That is, if Alice can devise a protocol that preserves the entanglement with another system, then this same protocol will also be able to preserve quantum information that she transmits.

The following theorem gives a simple way to represent the entanglement fidelity in terms of the Kraus operators of a given noisy quantum channel.

Theorem 9.5.1. *Given a quantum channel $\mathcal{N}$ with Kraus operators A_m, the entanglement fidelity $F_e(\rho,\mathcal{N})$ is equal to the following expression:*

$$F_e(\rho,\mathcal{N}) = \sum_m \left|\operatorname{Tr}\{\rho^A A_m\}\right|^2. \tag{9.195}$$

Proof. Suppose the spectral decomposition of ρ^A is $\rho^A = \sum_i \lambda_i |i\rangle\langle i|^A$. Its purification is then $|\psi\rangle^{RA} = \sum_i \sqrt{\lambda_i}|i\rangle^R|i\rangle^A$ where $\{|i\rangle^R\}$ is some orthonormal basis on the reference system. The entanglement fidelity is then as follows:

$$\begin{aligned}
&\langle\psi|\left(I^R\otimes\mathcal{N}^A\right)\left(\psi^{RA}\right)|\psi\rangle \\
&= \sum_{i,j,k,l,m} \sqrt{\lambda_i\lambda_j\lambda_k\lambda_l}\langle i|^R\langle i|^A\left(I^R\otimes A_m^A\right)|j\rangle^R|j\rangle^A\langle k|^R\langle k|^A\left(I^R\otimes A_m^{\dagger A}\right)|l\rangle^R|l\rangle^A && (9.196)\\
&= \sum_{i,j,k,l,m} \sqrt{\lambda_i\lambda_j\lambda_k\lambda_l}\langle i|j\rangle^R\langle i|^A A_m^A|j\rangle^A\langle k|l\rangle^R\langle k|^A A_m^{\dagger A}|l\rangle^A && (9.197)\\
&= \sum_{i,k,m} \lambda_i\lambda_k\langle i|^A A_m^A|i\rangle^A\langle k|^A A_m^{\dagger A}|k\rangle^A && (9.198)\\
&= \sum_{i,k,m} \lambda_i\operatorname{Tr}\left\{|i\rangle\langle i|^A A_m^A\right\}\lambda_k\operatorname{Tr}\left\{|k\rangle\langle k|^A A_m^{\dagger A}\right\} && (9.199)\\
&= \sum_m \operatorname{Tr}\{\rho^A A_m\}\operatorname{Tr}\{\rho^A A_m^\dagger\} && (9.200)\\
&= \sum_m \left|\operatorname{Tr}\{\rho^A A_m\}\right|^2. && (9.201)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

□

Exercise 9.5.1 Show that the entanglement fidelity is convex in the input state:

$$F_e(\lambda\rho_1 + (1-\lambda)\rho_2, \mathcal{N}) \leq \lambda F_e(\rho_1, \mathcal{N}) + (1-\lambda) F_e(\rho_2, \mathcal{N}). \tag{9.202}$$

(Hint: The result of Theorem 9.5.1 is useful here.)

9.5.3 Relationship between Expected Fidelity and Entanglement Fidelity

The entanglement fidelity and the expected fidelity provide seemingly different methods for quantifying the ability of a noisy quantum channel to preserve quantum information. Is there any way that we can show how they are related?

It turns out that they are indeed related. First, consider that the entanglement fidelity is a lower bound on the channel's fidelity for preserving the state ρ:

$$F_e(\rho, \mathcal{N}) \leq F(\rho, \mathcal{N}(\rho)). \tag{9.203}$$

The above result follows simply from the monotonicity of fidelity under partial trace (Lemma 9.2.1). We can show that the entanglement fidelity is always less than the expected fidelity in (9.190) by combining convexity of entanglement fidelity (Exercise 9.5.1) and the bound in (9.203):

$$\begin{aligned} F_e\left(\sum_x p_X(x)\rho_x, \mathcal{N}\right) &\leq \sum_x p_X(x) F_e(\rho_x, \mathcal{N}) && (9.204)\\ &\leq \sum_x p_X(x) F(\rho_x, \mathcal{N}(\rho_x)) && (9.205)\\ &= \overline{F}(\mathcal{N}) && (9.206) \end{aligned}$$

Thus, any channel that preserves entanglement with some reference system preserves the expected fidelity of an ensemble. In most cases, we only consider the entanglement fidelity as the defining measure of performance of a noisy quantum channel.

The relationship between entanglement fidelity and expected fidelity becomes more exact (and more beautiful) in the case where we select a random quantum state according to the Haar measure. It is possible to show that the expected fidelity in (9.192) relates to the entanglement fidelity as follows:

$$\overline{F}(\mathcal{N}) = \frac{dF_e(\pi, \mathcal{N}) + 1}{d+1}, \tag{9.207}$$

where d is the dimension of the input system and π is the maximally mixed state with purification to the maximally entangled state.

Exercise 9.5.2 Prove that the relation in (9.207) holds for a quantum depolarizing channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

9.6 The Hilbert-Schmidt Distance Measure

One final distance measure that we develop is the Hilbert-Schmidt distance measure. It is most similar to the familiar Euclidean distance measure of vectors because an ℓ_2-norm induces it. This distance measure does not have an appealing operational interpretation like the trace distance or fidelity do, and so we *do not* employ it to compare quantum states. Nevertheless, it can sometimes be helpful analytically to exploit this distance measure and to relate it to the trace distance via the bound in Exercise 9.6.1 below.

Let us define the Hilbert-Schmidt norm of an operator M as follows:

$$\|M\|_2 = \sqrt{\text{Tr}\{M^\dagger M\}}. \tag{9.208}$$

It is straightforward to show that the above norm meets the three requirements of a norm: positivity, homogeneity, and the triangle inequality. One can compute this norm simply by summing the squares of the eigenvalues of the operator M:

$$\|M\|_2 = \sum_x |\mu_x|^2, \tag{9.209}$$

where M admits the spectral decomposition $\sum_x \mu_x |x\rangle\langle x|$.

The Hilbert-Schmidt norm induces the following Hilbert-Schmidt distance measure:

$$\|M - N\|_2 = \sqrt{\text{Tr}\Big\{(M-N)^\dagger (M-N)\Big\}}. \tag{9.210}$$

We can of course then apply this distance measure to quantum states ρ and σ simply by plugging ρ and σ into the above formula in place of M and N.

The Hilbert-Schmidt distance measure sometimes finds use in the proofs of coding theorems in quantum Shannon theory because it is often easier to find good bounds on it rather than on the trace distance. In some cases, we might be taking expectations over ensembles of density operators and this expectation often reduces to computing variances or covariances.

Exercise 9.6.1 Show that the following inequality holds for any normal operator X

$$\|X\|_1^2 \leq d\|X\|_2^2, \tag{9.211}$$

where d is the dimension of the support of X. (Hint: use the convexity of the square function.)

9.7 History and Further Reading

Fuchs' thesis [96] and research paper [97] are a good starting point for learning more regarding trace distance and fidelity. Other notable sources are the book of Nielsen and Chuang [197], Yard's thesis [266], and Kretschmann's thesis [242]. Helstrom demonstrated the operational

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

interpretation of the trace distance in the context of quantum hypothesis testing [138, 139]. Uhlmann first proved his theorem in Ref. [240], and Jozsa later simplified the proof of this theorem in Ref. [165]. Schumacher introduced the entanglement fidelity in Ref. [217], and Barnum *et al.* made further observations regarding it in Ref. [15]. Nielsen provided a simple proof of the exact relation between entanglement fidelity and expected fidelity [196].

Winter originally proved the "Gentle Measurement" Lemma in Ref. [255] and in his thesis [256]. There, he used it to obtain a variation of the direct part of the HSW coding theorem. Later, he used it to prove achievable rates for the quantum multiple access channel [257]. Ogawa and Nagaoka subsequently improved this bound to $2\sqrt{\epsilon}$ in Appendix C of Ref. [199].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 10

Classical Information and Entropy

All physical systems register bits of information, whether it be an atom, an electrical current, the location of a billiard ball, or a switch. Information can be classical, quantum, or a hybrid of both, depending on the system. For example, an atom or an electron or a superconducting system can register *quantum* information because the quantum theory applies to each of these systems, but we can safely argue that the location of a billiard ball registers classical information only. These atoms or electrons or superconducting systems can also register classical bits because it is always possible for a quantum system to register classical bits.

The term *information*, in the context of information theory, has a precise meaning that is somewhat different from our prior "every day" experience with it. Recall that the notion of the physical bit refers to the physical representation of a bit, and the information bit is a measure of how much we learn from the outcome of a random experiment. Perhaps the word "surprise" better captures the notion of information as it applies in the context of information theory.

This chapter begins our formal study of classical information. Recall that Chapter 2 overviewed some of the major operational tasks in classical information theory. Here, our approach is somewhat different because our aim is to provide an intuitive understanding of information measures, in terms of the parties who have access to the classical systems. We define precise mathematical formulas that measure the amount of information encoded in a single physical system or in multiple physical systems. The advantage of developing this theory is that we can study information in its own right without having to consider the details of the physical system that registers it.

We first introduce the entropy in Section 10.1 as the expected surprise of a random variable. We extend this basic notion of entropy to develop other measures of information in Sections 10.2-10.6 that prove useful as intuitive informational measures, but also, and perhaps more importantly, these measures are the answers to operational tasks that one might wish to perform with noisy resources. While introducing these quantities, we discuss and prove several mathematical results concerning them that are important tools for the practicing information theorist. These tools are useful both for proving results and for increasing our understanding of the nature of information. Section 10.7 introduces informa-

tion inequalities that help us to understand the limits on our ability to process information. Section 10.8 ends the chapter by applying the classical informational measures developed in the forthcoming sections to the classical information that one can extract from a quantum system.

10.1 Entropy of a Random Variable

Consider a random variable X. Each realization x of random variable X belongs to an alphabet $\mathcal{X}$. Let $p_X(x)$ denote the probability density function of X so that $p_X(x)$ is the probability that realization x occurs. The information content $i(x)$ of a particular realization x is a measure of the surprise that one has upon learning the outcome of a random experiment:

$$i(x) \equiv -\log(p_X(x)). \tag{10.1}$$

The logarithm is base two and this choice implies that we measure surprise or information in bits.

Figure 10.1 plots the information content for values in the unit interval. This measure of surprise behaves as we would hope—it is higher for lower probability events that surprise us, and it is lower for higher probability events that do not surprise us. Inspection of the figure reveals that the information content is positive for any realization x.

The information content is also additive, due to the choice of the logarithm function. Given two independent random experiments involving random variable X with respective realizations x_1 and x_2, we have that

$$i(x_1, x_2) = -\log(p_{X,X}(x_1, x_2)) = -\log(p_X(x_1)p_X(x_2)) = i(x_1) + i(x_2). \tag{10.2}$$

Additivity is a property that we look for in measures of information (so much so that we dedicate the whole of Chapter 12 to this issue for more general measures of information).

The information content is a useful measure of surprise for particular realizations of random variable X, but it does not capture a general notion of the amount of surprise that a given random variable X possesses. The entropy $H(X)$ captures this general notion of the surprise of a random variable X—it is the expected information content of random variable X:

$$H(X) \equiv \mathbb{E}_X\{i(X)\}. \tag{10.3}$$

At a first glance, the above definition may seem strangely self-referential because the argument of the probability density function $p_X(x)$ is itself the random variable X, but this is well-defined mathematically. Evaluating the above formula gives the following expression for the entropy $H(X)$:

$$H(X) \equiv -\sum_x p_X(x)\log(p_X(x)). \tag{10.4}$$

We adopt the convention that $0 \cdot \log(0) = 0$ for realizations with zero probability. The fact that $\lim_{\epsilon\to 0} \epsilon \cdot \log(\epsilon) = 0$ intuitively justifies this latter convention.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

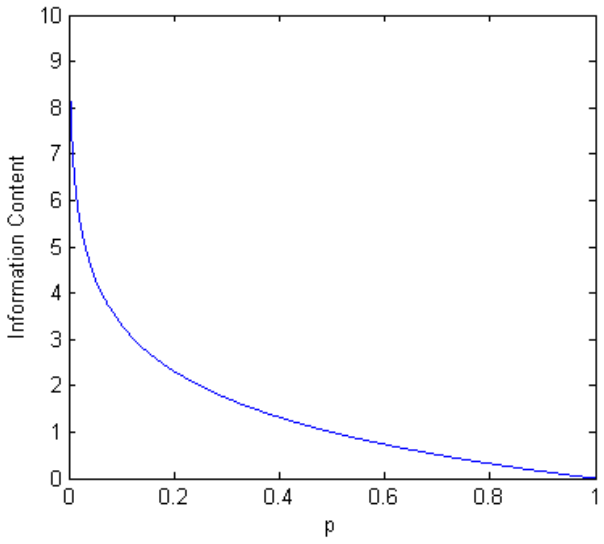

Figure 10.1: The information content or "surprise" in (10.1) as a function of a probability p ranging from 0 to 1. An event has a lower surprise if it is more likely to occur and it has a higher surprise if it less likely to occur.

The entropy admits an intuitive interpretation. Suppose that Alice generates a random experiment in her lab that selects a realization x according to the density $p_X(x)$ of random variable X. Suppose further that Bob has not yet learned the outcome of the experiment. The interpretation of the entropy $H(X)$ is that it quantifies Bob's uncertainty about X before learning it—his expected information gain is $H(X)$ bits upon learning the outcome of the random experiment. Shannon's noiseless coding theorem, described in Chapter 2, makes this interpretation precise by proving that Alice needs to send Bob bits at a rate $H(X)$ in order for him to be able to decode a compressed message. Figure 10.2(a) depicts the interpretation of the entropy $H(X)$, along with a similar interpretation for the conditional entropy that we introduce in Section 10.2.

10.1.1 The Binary Entropy Function

A special case of the entropy occurs when the random variable X is a Bernoulli random variable with probability density $p_X(0) = p$ and $p_X(1) = 1 - p$. This Bernoulli random variable could correspond to the outcome of a random coin flip. The entropy in this case is known as the *binary entropy function*:

$$H(p) \equiv -p \log p - (1-p) \log(1-p). \tag{10.5}$$

It quantifies the number of bits that we learn from the outcome of the coin flip. If the coin is unbiased ($p = 1/2$), then we learn a maximum of one bit ($H(p) = 1$). If the coin is deterministic ($p = 0$ or $p = 1$), then we do not learn anything from the outcome ($H(p) = 0$).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

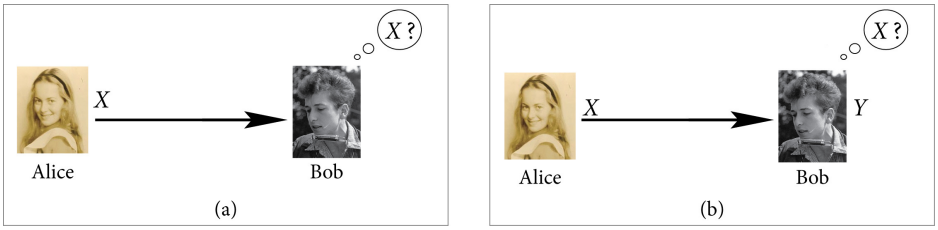

Figure 10.2: (a) The entropy $H(X)$ is the uncertainty that Bob has about random variable X before learning it. (b) The conditional entropy $H(X|Y)$ is the uncertainty that Bob has about X when he already possesses Y.

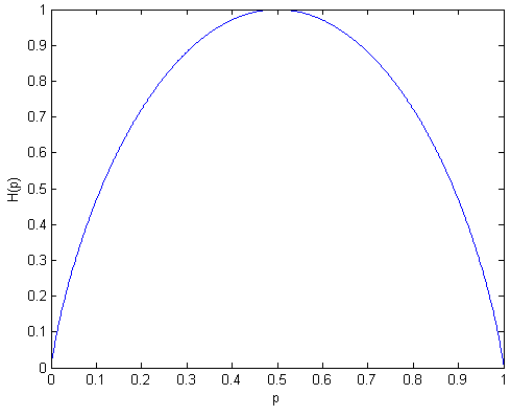

Figure 10.3: The binary entropy function $H(p)$ displayed as a function of the parameter p.

Figure 10.3 displays a plot of the binary entropy function. The figure reveals that the binary entropy function $H(p)$ is a concave function of the parameter p and has its peak at $p = 1/2$.

10.1.2 Mathematical Properties of Entropy

We now discuss five important mathematical properties of the entropy $H(X)$.

Property 10.1.1 (Positivity) The entropy $H(X)$ is non-negative for any probability density $p_X(x)$:

$$H(X) \geq 0. \tag{10.6}$$

Proof. Positivity follows because entropy is the expected information content $i(x)$, and the information content itself is positive. It is perhaps intuitive that the entropy should be positive because positivity implies that we always learn some number of bits upon learning random variable X (if we already know beforehand what the outcome of a random experiment will be, then we learn zero bits of information once we perform it). In a classical sense, we can never learn a negative amount of information! □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Property 10.1.2 (Concavity) The entropy $H(X)$ is concave in the probability density $p_X(x)$.

Proof. We justify this result with a heuristic "mixing" argument for now, and provide a formal proof in Section 10.7.1. Consider two random variables X_1 and X_2 with two respective probability density functions $p_{X_1}(x)$ and $p_{X_2}(x)$ whose realizations belong to the same alphabet. Consider a Bernoulli random variable B with probabilities q and $1-q$ corresponding to its two respective realizations $b = 1$ and $b = 2$. Suppose that we first generate a realization b of random variable B and then generate a realization x of random variable X_b. Random variable X_B then denotes a mixed version of the two random variables X_1 and X_2. The probability density of X_B is $p_{X_B}(x) = qp_{X_1}(x) + (1-q)p_{X_2}(x)$. Concavity of entropy is the following inequality:

$$H(X_B) \geq qH(X_1) + (1-q)H(X_2). \tag{10.7}$$

Our heuristic argument is that this mixing process leads to more uncertainty for the mixed random variable X_B than the expected uncertainty over the two individual random variables. We can think of this result as a physical situation involving two gases. Two gases each have their own entropy, but the entropy increases when we mix the two gases together. We later give a more formal argument to justify concavity. □

Property 10.1.3 (Invariance under permutations) The entropy is invariant under permutations of the realizations of random variable X.

Proof. That is, suppose that we apply some permutation π to realizations $x_1, x_2, \ldots, x_{|\mathcal{X}|}$ so that they respectively become $\pi(x_1), \pi(x_2), \ldots, \pi(x_{|\mathcal{X}|})$. Then the entropy is invariant under this shuffling because it depends only on the probabilities of the realizations, not the values of the realizations. □

Property 10.1.4 (Minimum value) The entropy vanishes for a deterministic variable.

Proof. We would expect that the entropy of a *deterministic* variable should vanish, given the interpretation of entropy as the uncertainty of a random experiment. This intuition holds true and it is the degenerate probability density $p_X(x) = \delta_{x,x_0}$, where the realization x_0 has all the probability and other realizations have vanishing probability, that gives the minimum value of the entropy: $H(X) = 0$ when X has a degenerate density. □

Sometimes, we may not have any prior information about the possible values of a variable in a system, and we may decide that it is most appropriate to describe them with a probability density function. How should we assign this probability density if we do not have any prior information about the values? Theorists and experimentalists often resort to a "principle of maximum entropy" or a "principle of maximal ignorance"—we should assign the probability density to be the one that maximizes the entropy.

Property 10.1.5 (Maximum value) The maximum value of the entropy $H(X)$ for a random variable X with d different realizations is $\log d$:

$$H(X) \leq \log d. \tag{10.8}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Proof. For any variable with a finite number of values, the probability density that maximizes the entropy is the uniform distribution. This distribution results in an entropy $\log d$, where d is the number of values for the variable (the result of Exercise 2.1.1 is that $\log d$ is the entropy of the uniform random variable). We can prove the above inequality with a simple Lagrangian optimization by solving for the density $p_X(x)$ that maximizes the entropy. Lagrangian optimization is well-suited for this task because the entropy is concave in the probability density, and thus any local maximum will be a global maximum. The Lagrangian $\mathcal{L}$ is as follows:

$$\mathcal{L} \equiv H(X) + \lambda\left(\sum_x p_X(x) - 1\right), \tag{10.9}$$

where $H(X)$ is the quantity that we are maximizing, subject to the constraint that the probability density $p_X(x)$ sums to unity. The partial derivative $\frac{\partial \mathcal{L}}{\partial p_X(x)}$ is as follows:

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial p_X(x)} \\
&= \frac{\partial}{\partial p_X(x)}\left(-\sum_{x'} p_X(x')\log(p_X(x')) + \lambda\left(\sum_{x'} p_X(x') - 1\right)\right) &(10.10)\\
&= -\log(p_X(x)) - 1 + \lambda &(10.11)
\end{aligned}$$

We null the partial derivative $\frac{\partial \mathcal{L}}{\partial p_X(x)}$ to find the density that maximizes $\mathcal{L}$:

$$\begin{aligned}
0 &= -\log(p_X(x)) - 1 + \lambda &(10.12)\\
\Rightarrow p_X(x) &= 2^{\lambda - 1}. &(10.13)
\end{aligned}$$

The resulting density $p_X(x)$ is dependent only on a constant λ, implying that it must be uniform $p_X(x) = \frac{1}{d}$. Thus, the uniform distribution $\frac{1}{d}$ maximizes the entropy $H(X)$ when random variable X is finite. □

10.2 Conditional Entropy

Let us now suppose that Alice possesses random variable X and Bob possesses some other random variable Y. Random variables X and Y share correlations if they are not statistically independent, and Bob then possesses "side information" about X in the form of Y. Let $i(x|y)$ denote the conditional information content:

$$i(x|y) \equiv -\log\big(p_{X|Y}(x|y)\big). \tag{10.14}$$

The entropy $H(X|Y=y)$ of random variable X conditional on a particular realization y of random variable Y is the expected conditional information content, where the expectation is with respect to X:

$$\begin{aligned}
H(X|Y=y) &\equiv \mathbb{E}_X\{i(X|y)\} &(10.15)\\
&= -\sum_x p_{X|Y}(x|y)\log\big(p_{X|Y}(x|y)\big). &(10.16)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The relevant entropy that applies to the scenario where Bob possesses side information is the conditional entropy $H(X|Y)$. It is the expected conditional information content where the expectation is with respect to both X and Y:

$$\begin{aligned} H(X|Y) &\equiv \mathbb{E}_{X,Y}\{i(X|Y)\} && (10.17)\\ &= \sum_y p_Y(y) H(X|Y=y) && (10.18)\\ &= -\sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \log\big(p_{X|Y}(x|y)\big) && (10.19)\\ &= -\sum_{x,y} p_{X,Y}(x,y) \log\big(p_{X|Y}(x|y)\big). && (10.20) \end{aligned}$$

The conditional entropy $H(X|Y)$ as well deserves an interpretation. Suppose that Alice possesses random variable X and Bob possesses random variable Y. The conditional entropy $H(X|Y)$ is the amount of uncertainty that Bob has about X given that he already possesses Y. Figure 10.2(b) depicts this interpretation.

The above interpretation of the conditional entropy $H(X|Y)$ immediately suggests that it should be less than or equal to the entropy $H(X)$. That is, having access to a side variable Y should only decrease our uncertainty about another variable. We state this idea as the following theorem and give a formal proof in Section 10.7.1.

Theorem 10.2.1 (Conditioning does not increase entropy)**.** *The entropy $H(X)$ is greater than or equal to the conditional entropy $H(X|Y)$:*

$$H(X) \geq H(X|Y). \tag{10.21}$$

Positivity of conditional entropy follows from positivity of entropy because conditional entropy is the expectation of the entropy $H(X|Y=y)$ with respect to the density $p_Y(y)$. It is again intuitive that conditional entropy should be positive. Even if we have access to some side information Y, we always learn some number of bits of information upon learning the outcome of a random experiment involving X. Perhaps strangely, we will see that *quantum* conditional entropy can become negative, defying our intuition of information in the classical sense given here.

10.3 Joint Entropy

What if Bob knows neither X nor Y? The natural entropic quantity that describes his uncertainty is the joint entropy $H(X,Y)$. The joint entropy is merely the entropy of the joint random variable (X,Y):

$$\begin{aligned} H(X,Y) &\equiv \mathbb{E}_{X,Y}\{i(X,Y)\} && (10.22)\\ &= -\sum_{x,y} p_{X,Y}(x,y) \log(p_{X,Y}(x,y)). && (10.23) \end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The following exercise asks you to explore the relation between joint entropy $H(X,Y)$, conditional entropy $H(Y|X)$, and marginal entropy $H(X)$. Its proof follows by considering that the multiplicative probability relation $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$ of joint probability, conditional probability, and marginal entropy becomes an additive relation under the logarithms of the entropic definitions.

Exercise 10.3.1 Verify that $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.

Exercise 10.3.2 Extend the result of Exercise 10.3.1 to prove the following chaining rule for entropy:

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1}, \ldots, X_1). \tag{10.24}$$

Exercise 10.3.3 Prove that entropy is *subadditive*:

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i), \tag{10.25}$$

by exploiting Theorem 10.2.1 and the entropy chaining rule in Exercise 10.3.2.

Exercise 10.3.4 Prove that entropy is additive when the random variables $X_1, \ldots, X_n$ are independent:

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i). \tag{10.26}$$

10.4 Mutual Information

We now introduce an entropic measure of the common or mutual information that two parties possess. Suppose that Alice possesses random variable X and Bob possesses random variable Y. The mutual information is the marginal entropy $H(X)$ less the conditional entropy $H(X|Y)$:

$$I(X;Y) \equiv H(X) - H(X|Y). \tag{10.27}$$

It quantifies the dependence or correlations of the two random variables X and Y.

The mutual information measures how much knowing one random variable reduces the uncertainty about the other random variable. In this sense, it is the common information between the two random variables. Bob possesses Y and thus has an uncertainty $H(X|Y)$ about Alice's variable X. Knowledge of Y gives an information gain of $H(X|Y)$ bits about X and then reduces the overall uncertainty $H(X)$ about X, the uncertainty were he not to have any side information at all about X.

Exercise 10.4.1 Show that the mutual information is symmetric in its inputs:

$$I(X;Y) = I(Y;X), \tag{10.28}$$

implying additionally that

$$I(X;Y) = H(Y) - H(Y|X). \tag{10.29}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can also express the mutual information $I(X;Y)$ in terms of the respective joint and marginal probability density functions $p_{X,Y}(x,y)$ and $p_X(x)$ and $p_Y(y)$:

$$I(X;Y) = \sum_{x,y} p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right). \tag{10.30}$$

The above expression leads to two insights regarding the mutual information $I(X;Y)$. Two random variables X and Y possess zero bits of mutual information if and only if they are statistically independent (recall that the joint density factors as $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ when X and Y are independent). That is, knowledge of Y does not give any information about X when the random variables are statistically independent. Also, two random variables possess $H(X)$ bits of mutual information if they are perfectly correlated in the sense that $Y = X$.

Theorem 10.4.1 below states that the mutual information $I(X;Y)$ is non-negative for any random variables X and Y—we provide a formal proof in Section 10.7.1. Though, this follows naturally from the definition of mutual information in (10.27) and "conditioning does not increase entropy" (Theorem 10.2.1).

Theorem 10.4.1. *The mutual information $I(X;Y)$ is non-negative for any random variables X and Y:*

$$I(X;Y) \geq 0. \tag{10.31}$$

10.5 Relative Entropy

The relative entropy is another important entropic quantity that quantifies how "far" one probability density function $p_{X_1}(x)$ is from another probability density function $p_{X_2}(x)$. We define the relative entropy $D(p_{X_1}||p_{X_2})$ as follows:

$$D(p_{X_1}||p_{X_2}) \equiv \sum_x p_{X_1}(x) \log\left(\frac{p_{X_1}(x)}{p_{X_2}(x)}\right). \tag{10.32}$$

According to the above definition, the relative entropy is an expected log-likelihood ratio of the densities $p_{X_1}(x)$ and $p_{X_2}(x)$.

The above definition implies that the relative entropy is not symmetric under interchange of the densities $p_{X_1}(x)$ and $p_{X_2}(x)$. Thus, the relative entropy is not a distance measure in the strict mathematical sense because it is not symmetric (nor does it satisfy a triangle inequality).

The relative entropy has an interpretation in source coding. Suppose that an information source generates a random variable X_1 according to the density $p_{X_1}(x)$. Suppose further that Alice (the compressor) mistakenly assumes that the probability density of the information source is instead $p_{X_2}(x)$ and codes according to this density. Then the relative entropy quantifies the inefficiency that Alice incurs when she codes according to the mistaken probability

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

density—Alice requires $H(X_1) + D(p_{X_1}||p_{X_2})$ bits on average to code (whereas she would only require $H(X_1)$ bits on average to code if she used the true density $p_{X_1}(x)$).

We might also see now that the mutual information $I(X;Y)$ is equivalent to the relative entropy $D(p_{X,Y}(x,y)||p_X(x)p_Y(y))$ by comparing the definition of relative entropy in (10.32) and the expression for the mutual information in (10.30). In this sense, the mutual information quantifies how far the two random variables X and Y are from being independent because it calculates the distance of the joint density $p_{X,Y}(x,y)$ from the product of the marginals $p_X(x)p_Y(y)$.

The relative entropy $D(p_{X_1}||p_{X_2})$ admits a pathological property. It can become infinite if the distribution $p_{X_1}(x_1)$ does not have all of its support contained in the support of $p_{X_2}(x_2)$ (i.e., if there is some realization x for which $p_{X_1}(x) \neq 0$ but $p_{X_2}(x) = 0$). This can be somewhat bothersome if we like this interpretation of relative entropy as a notion of distance. In an extreme case, we would think that the distance between a deterministic binary random variable X_2 where $\Pr\{X_2 = 1\} = 1$ and one with probabilities $\Pr\{X_1 = 0\} = \epsilon$ and $\Pr\{X_1 = 1\} = 1 - \epsilon$ should be on the order of ϵ (this is true for the Komolgorov distance). Though, the relative entropy $D(p_{X_1}||p_{X_2})$ in this case is infinite, in spite of our intuition that these distributions are close. The interpretation in lossless source coding is that it would require an infinite number of bits to code a distribution p_{X_1} losslessly if Alice mistakes it as p_{X_2}. Alice thinks that the symbol $X_2 = 0$ never occurs, and in fact, she thinks that the typical set consists of just one sequence of all ones and every other sequence is atypical. But in reality, the typical set is quite a bit larger than this, and it is only in the limit of an infinite number of bits that we can say her compression is truly lossless.

10.6 Conditional Mutual Information

What is the common information between two random variables X and Y when we have some side information embodied in a random variable Z? The entropic quantity that answers this question is the conditional mutual information. It is simply the mutual information conditioned on a random variable Z:

$$\begin{aligned} I(X;Y|Z) &\equiv H(Y|Z) - H(Y|X,Z) && (10.33)\\ &= H(X|Z) - H(X|Y,Z) && (10.34)\\ &= H(X|Z) + H(Y|Z) - H(X,Y|Z). && (10.35) \end{aligned}$$

Theorem 10.6.1 (Strong subadditivity)**.** *The conditional mutual information $I(X;Y|Z)$ is non-negative:*

$$I(X;Y|Z) \geq 0. \tag{10.36}$$

Proof. The proof of the above theorem is a straightforward consequence of the positivity of mutual information (Theorem 10.4.1). Consider the following equivalence:

$$I(X;Y|Z) = \sum_z p_Z(z) I(X;Y|Z=z), \tag{10.37}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $I(X;Y|Z=z)$ is a mutual information with respect to the joint density $p_{X,Y|Z}(x,y|z)$ and the marginal densities $p_{X|Z}(x|z)$ and $p_{Y|Z}(y|z)$. Positivity of $I(X;Y|Z)$ then follows from positivity of $p_Z(z)$ and $I(X;Y|Z=z)$. □

The proof of the above classical version of strong subadditivity is perhaps trivial in hindsight (it requires only a few arguments). The proof of the quantum version of strong subaddivity is highly nontrivial on the other hand. We discuss strong subadditivity of quantum entropy in the next chapter.

Theorem 10.6.2. *The conditional mutual information vanishes if random variables X and Y are conditionally independent through Z. That is,*

$$I(X;Y|Z) = 0, \tag{10.38}$$

if $p_{X,Y|Z}(x,y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$.

Proof. We can establish the proof by expressing the conditional mutual information in a form similar to that for the mutual information in (10.30):

$$I(X;Y|Z) = \sum_{x,y} p_{X,Y|Z}(x,y|z) \log\left(\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)}\right). \tag{10.39}$$

The logarithm then vanishes when $p_{X,Y|Z}(x,y|z)$ factors as $p_{X|Z}(x|z)p_{Y|Z}(y|z)$. □

Exercise 10.6.1 The expression in (10.36) represents the most compact way to express the strong subadditivity of entropy. Show that the following inequalities are equivalent ways of representing strong subadditivity:

$$H(XY|Z) \leq H(X|Z) + H(Y|Z), \tag{10.40}$$
$$H(XYZ) + H(Z) \leq H(XZ) + H(YZ), \tag{10.41}$$
$$H(X|YZ) \leq H(X|Z). \tag{10.42}$$

Exercise 10.6.2 Prove the following chaining rule for mutual information:

$$I(X_1,\ldots,X_n;Y) = I(X_1;Y) + I(X_2;Y|X_1) + \cdots + I(X_n;Y|X_1,\ldots X_{n-1}). \tag{10.43}$$

10.7 Information Inequalities

The entropic quantities introduced in the previous sections each have bounds associated with them. These bounds are fundamental limits on our ability to process and store information. We introduce three bounds in this section: the fundamental information inequality, the data processing inequality, and Fano's inequality. Each of these inequalities plays an important role in information theory, and we describe these roles in more detail in the forthcoming subsections.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

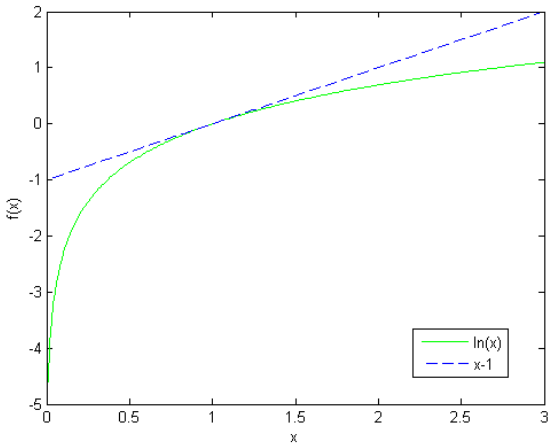

Figure 10.4: A plot that compares the functions $\ln x$ and $x-1$, revealing that $\ln x \leq x-1$ for all positive x.

10.7.1 The Fundamental Information Inequality

The *fundamental information inequality* is the statement that the relative entropy is always non-negative. This seemingly innocuous result has several important implications—namely, the maximal value of entropy, conditioning does not increase entropy (Theorem 10.2.1), positivity of mutual information (Theorem 10.4.1), and strong subadditivity (Theorem 10.6.1) are straightforward corollaries of it. The proof of the fundamental information inequality follows from the application of a simple inequality: $\ln x \leq x - 1$.

Theorem 10.7.1 (Positivity of Relative Entropy)**.** *The relative entropy $D(p_{X_1}||p_{X_2})$ is non-negative for any probability density functions $p_{X_1}(x)$ and $p_{X_2}(x)$:*

$$D(p_{X_1}||p_{X_2}) \geq 0. \tag{10.44}$$

Proof. The proof relies on the inequality $\ln x \leq x - 1$ that holds for all positive x and saturates for $x = 1$. Figure 10.4 plots these functions. We prove the theorem by application of the following chain of inequalities:

$$D(p_{X_1}||p_{X_2}) = \sum_x p_{X_1}(x) \log\left(\frac{p_{X_1}(x)}{p_{X_2}(x)}\right) \tag{10.45}$$

$$= -\frac{1}{\ln 2} \sum_x p_{X_1}(x) \ln\left(\frac{p_{X_2}(x)}{p_{X_1}(x)}\right) \tag{10.46}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$$\geq \frac{1}{\ln 2} \sum_x p_{X_1}(x) \left(1 - \frac{p_{X_2}(x)}{p_{X_1}(x)}\right) \tag{10.47}$$

$$= \frac{1}{\ln 2} \left(\sum_x p_{X_1}(x) - \sum_x p_{X_2}(x)\right) \tag{10.48}$$

$$= 0. \tag{10.49}$$

The sole inequality follows because $-\ln x \geq 1-x$ (a simple rearrangement of $\ln x \leq x-1$). $\square$

We can now quickly prove several corollaries of the above theorem. Recall in Section 10.1.2 that we proved that the entropy $H(X)$ takes the maximal value $\log d$, where d is size of the alphabet of X. The proof method involved Lagrange multipliers. Here, we can prove this result simply by computing the relative entropy $D\big(p_X(x)||\frac{1}{d}\big)$, where $p_X(x)$ is the probability density of X and $\frac{1}{d}$ is the uniform density, and applying the fundamental information inequality:

$$0 \leq D\left(p_X(x)||\frac{1}{d}\right) \tag{10.50}$$

$$= \sum_x p_X(x) \log \left(\frac{p_X(x)}{\frac{1}{d}}\right) \tag{10.51}$$

$$= -H(X) + \sum_x p_X(x) \log d \tag{10.52}$$

$$= -H(X) + \log d. \tag{10.53}$$

It then follows that $H(X) \leq \log d$ by combining the first line with the last.

Positivity of mutual information (Theorem 10.4.1) follows by recalling that

$$I(X;Y) = D(p_{X,Y}(x,y)||p_X(x)p_Y(y)) \tag{10.54}$$

and applying the fundamental information inequality. Conditioning does not increase entropy (Theorem 10.2.1) follows by noting that $I(X;Y) = H(X) - H(X|Y)$ and applying Theorem 10.4.1.

10.7.2 Data Processing Inequality

Another important inequality in classical information theory is the *data processing inequality*. This inequality states that correlations between random variables can only decrease after we process one variable according to some stochastic function that depends only on that variable. The data processing inequality finds application in the converse proof of a coding theorem (the proof of the optimality of a communication rate).

We detail the scenario that applies for the data processing inequality. Suppose that we initially have two random variables X and Y. We might say that random variable Y arises from random variable X by processing X according to a stochastic map $\mathcal{N}_1 \equiv$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

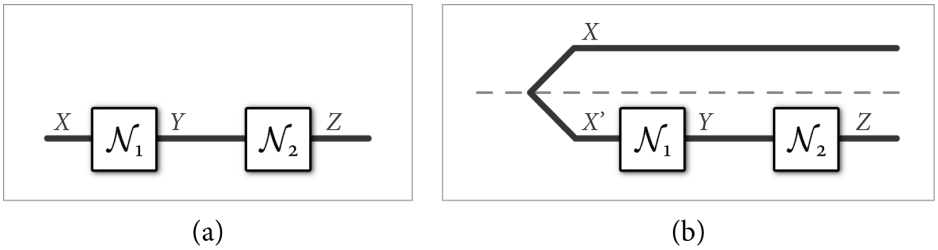

Figure 10.5: Two slightly different depictions of the scenario in the data processing inequality. (a) The map $\mathcal{N}_1$ processes random variable X to produce some random variable Y, and the map $\mathcal{N}_2$ processes the random variable Y to produce the random variable Z. The inequality $I(X;Y) \geq I(X;Z)$ applies here because correlations can only decrease after data processing. (b) This depiction of data processing helps us to build intuition for data processing in the quantum world. The protocol begins with two perfectly correlated random variables X and X'—perfect correlation implies that $p_{X,X'}(x,x') = p_X(x)\delta_{x,x'}$ and further that $H(X) = I(X;X')$. We process random variable X' with a stochastic map $\mathcal{N}_1$ to produce a random variable Y, and then further process Y according to the stochastic map $\mathcal{N}_2$ to produce random variable Z. By the data processing inequality, the following chain of inequalities holds: $I(X;X') \geq I(X;Y) \geq I(X;Z)$.

$p_{Y|X}(y|x)$. That is, the two random variables arise by first picking X according to the density $p_X(x)$ and then processing X according to the stochastic map $\mathcal{N}_1$. The mutual information $I(X;Y)$ quantifies the correlations between these two random variables. Suppose then that we process Y according to some other stochastic map $\mathcal{N}_2 \equiv p_{Z|Y}(z|y)$ to produce a random variable Z (note that the map can also be deterministic because the set of stochastic maps subsumes the set of deterministic maps). Then the data processing inequality states that the correlations between X and Z must be less than the correlations between X and Y:

$$I(X;Y) \geq I(X;Z), \tag{10.55}$$

because data processing according to any map $\mathcal{N}_2$ can only decrease correlations. Figure 10.5(a) depicts the scenario described above. Figure 10.5(b) depicts a slightly different scenario for data processing that helps build intuition for the forthcoming notion of quantum data processing in Section 11.9.3 of the next chapter. Theorem 10.7.2 below states the classical data processing inequality.

The scenario described in the above paragraph contains a major assumption and you may have picked up on it. We assumed that the stochastic map $p_{Z|Y}(z|y)$ that produces random variable Z depends on random variable Y only—it has no dependence on X. It then holds that

$$p_{Z|Y,X}(z|y,x) = p_{Z|Y}(z|y). \tag{10.56}$$

This assumption is called the Markovian assumption and is the crucial assumption in the proof of the data processing inequality. We say that the three random variables X, Y, and Z form a *Markov chain* and use the notation $X \to Y \to Z$ to indicate this stochastic relationship.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Theorem 10.7.2 (Data Processing Inequality)**.** *Suppose three random variables X, Y, and Z form a Markov chain: $X \to Y \to Z$. Then the following data processing inequality applies*

$$I(X;Y) \geq I(X;Z). \tag{10.57}$$

Proof. The Markov condition $X \to Y \to Z$ implies that random variables X and Z are conditionally independent through Y because

$$\begin{aligned} p_{X,Z|Y}(x,z|y) &= p_{Z|Y,X}(z|y,x)p_{X|Y}(x|y) && (10.58)\\ &= p_{Z|Y}(z|y)p_{X|Y}(x|y). && (10.59)\end{aligned}$$

We prove the data processing inequality by manipulating the mutual information $I(X;YZ)$. Consider the following equalities:

$$\begin{aligned} I(X;YZ) &= I(X;Y) + I(X;Z|Y) && (10.60)\\ &= I(X;Y). && (10.61)\end{aligned}$$

The first equality follows from the chain rule for mutual information (Exercise 10.6.2). The second equality follows because the conditional mutual information $I(X;Z|Y)$ vanishes for a Markov chain $X \to Y \to Z$—i.e., X and Z are conditionally independent through Y (recall Theorem 10.6.2). We can also expand the mutual information $I(X;YZ)$ in another way to obtain

$$I(X;YZ) = I(X;Z) + I(X;Y|Z). \tag{10.62}$$

Then the following equality holds for a Markov chain $X \to Y \to Z$ by exploiting (10.61):

$$I(X;Y) = I(X;Z) + I(X;Y|Z). \tag{10.63}$$

The inequality in Theorem 10.7.2 follows because $I(X;Y|Z)$ is non-negative for any random variables X, Y, and Z (recall Theorem 10.6.1). □

Corollary 10.7.1. *The following inequality holds for a Markov chain $X \to Y \to Z$:*

$$I(X;Y) \geq I(X;Y|Z). \tag{10.64}$$

Proof. The proof follows by inspecting the above proof. □

10.7.3 Fano's Inequality

The last classical information inequality that we consider is Fano's inequality. This inequality also finds application in the converse proof of a coding theorem.

Fano's inequality applies to a general classical communication scenario. Suppose Alice possesses some random variable X that she transmits to Bob over a noisy communication channel. Let $p_{Y|X}(y|x)$ denote the stochastic map corresponding to the noisy communication channel. Bob receives a random variable Y from the channel and processes it in some way

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

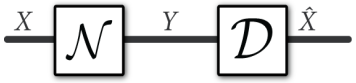

Figure 10.6: The classical communication scenario relevant in Fano's inequality. Alice transmits a random variable X over a noisy channel $\mathcal{N}$, producing a random variable Y. Bob receives Y and processes it according to some decoding map $\mathcal{D}$ to produce his best estimate $\hat{X}$ of Y.

to produce his best estimate $\hat{X}$ of the original random variable X. Figure 10.6 depicts this scenario.

The natural performance metric of this communication scenario is the probability of error $p_e \equiv \Pr\left\{\hat{X} \neq X\right\}$—a low probability of error corresponds to good performance. On the other hand, consider the conditional entropy $H(X|Y)$. We interpreted it before as the uncertainty about X from the perspective of someone who already knows Y. If the channel is noiseless ($p_{Y|X}(y|x) = \delta_{y,x}$), then there is no uncertainty about X because Y is identical to X:

$$H(X|Y) = 0. \tag{10.65}$$

As the channel becomes noisier, the conditional entropy $H(X|Y)$ increases away from zero. In this sense, the conditional entropy $H(X|Y)$ quantifies the information about X that is lost in the channel noise. We then might naturally expect there to be a relationship between the probability of error p_e and the conditional entropy $H(X|Y)$: the amount of information lost in the channel should be low if the probability of error is low. Fano's inequality provides a quantitative bound corresponding to this idea.

Theorem 10.7.3 (Fano's Inequality)**.** *Suppose that Alice sends a random variable X through a noisy channel to produce random variable Y and further processing of Y gives an estimate $\hat{X}$ of X. Thus, $X \to Y \to \hat{X}$ forms a Markov chain. Let $p_e \equiv \Pr\left\{\hat{X} \neq X\right\}$ denote the probability of error. Then the following function of the error probability p_e bounds the information lost in the channel noise:*

$$H(X|Y) \leq H_2(p_e) + p_e \log(|\mathcal{X}| - 1), \tag{10.66}$$

where $H_2(p_e)$ is the binary entropy function. In particular, note that

$$\lim_{p_e \to 0} H_2(p_e) + p_e \log(|\mathcal{X}| - 1) = 0. \tag{10.67}$$

Proof. Let E denote an indicator random variable that indicates whether an error occurs:

$$E = \begin{cases} 0 & : \quad X = \hat{X} \\ 1 & : \quad X \neq \hat{X} \end{cases}. \tag{10.68}$$

Consider the entropy

$$H\left(EX|\hat{X}\right) = H\left(X|\hat{X}\right) + H\left(E|X\hat{X}\right). \tag{10.69}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The entropy $H\left(E|X\hat{X}\right)$ on the RHS vanishes because there is no uncertainty about the indicator random variable E if we know both X and $\hat{X}$. Thus,

$$H\left(EX|\hat{X}\right) = H\left(X|\hat{X}\right). \tag{10.70}$$

Also, the data processing inequality applies to the Markov chain $X \to Y \to \hat{X}$:

$$I(X;Y) \geq I\left(X;\hat{X}\right), \tag{10.71}$$

and implies the following inequality:

$$H\left(X|\hat{X}\right) \geq H(X|Y). \tag{10.72}$$

Consider the following chain of inequalities:

$$\begin{aligned} H\left(EX|\hat{X}\right) &= H\left(E|\hat{X}\right) + H\left(X|E\hat{X}\right) && (10.73)\\ &\leq H(E) + H\left(X|E\hat{X}\right) && (10.74)\\ &= H_2(p_e) + p_e H\left(X|\hat{X}, E=1\right) \\ &\qquad + (1-p_e) H\left(X|\hat{X}, E=0\right) && (10.75)\\ &\leq H_2(p_e) + p_e \log(|\mathcal{X}|-1). && (10.76) \end{aligned}$$

The first equality follows by expanding the entropy $H\left(EX|\hat{X}\right)$. The first inequality follows because conditioning reduces entropy. The second equality follows by explicitly expanding the conditional entropy $H\left(X|E\hat{X}\right)$ in terms of the two possibilities of the error random variable E. The last inequality follows from two facts: there is no uncertainty about X when there is no error (when $E=0$) and $\hat{X}$ is available, and the uncertainty about X when there is an error (when $E=1$) and $\hat{X}$ is available is less than the uncertainty of a uniform distribution $\frac{1}{|\mathcal{X}|-1}$ on all of the other possibilities. Fano's inequality follows from putting together (10.72), (10.70), and (10.76):

$$H(X|Y) \leq H\left(X|\hat{X}\right) = H\left(EX|\hat{X}\right) \leq H_2(p_e) + p_e \log(|\mathcal{X}|-1). \tag{10.77}$$

□

10.8 Classical Information and Entropy of Quantum Systems

We can always process classical information by employing a quantum system as the carrier of information. The inputs and the outputs to a quantum protocol can both be classical.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

For example, we can prepare a quantum state according to some random variable X—the ensemble $\{p_X(x), \rho_x\}$ captures this idea. We can retrieve classical information from a quantum state in the form of some random variable Y by performing a measurement—the POVM $\{\Lambda_y\}$ captures this notion (recall that we employ the POVM formalism from Section 4.2.1 if we do not care about the state after the measurement). Suppose that Alice prepares a quantum state according to the ensemble $\{p_X(x), \rho_x\}$ and Bob measures the state according to the POVM $\{\Lambda_y\}$. Recall that the following formula gives the conditional probability $p_{Y|X}(y|x)$:

$$p_{Y|X}(y|x) = \text{Tr}\{\Lambda_y \rho_x\}. \tag{10.78}$$

Is there any benefit to processing classical information with quantum systems? Later, in Chapter 19, we see that there indeed is an enhanced performance because we can achieve higher communication rates in general by processing classical data with quantum resources. For now, we extend our notions of entropy in a straightforward way to include the above ideas.

10.8.1 Shannon Entropy of a POVM

The first notion that we can extend is the Shannon entropy, by determining the Shannon entropy of a POVM. Suppose that Alice prepares a quantum state ρ (there is no classical index here). Bob can then perform a particular POVM $\{\Lambda_x\}$ to learn about the quantum system. Let X denote the random variable corresponding to the classical output of the POVM. The probability density function $p_X(x)$ of random variable X is then

$$p_X(x) = \text{Tr}\{\Lambda_x \rho\}. \tag{10.79}$$

The Shannon entropy $H(X)$ of the POVM $\{\Lambda_x\}$ is

$$H(X) = -\sum_x p_X(x) \log(p_X(x)) \tag{10.80}$$

$$= -\sum_x \text{Tr}\{\Lambda_x \rho\} \log(\text{Tr}\{\Lambda_x \rho\}). \tag{10.81}$$

In the next chapter, we prove that the minimum Shannon entropy over all rank-one POVMs is equal to a quantity known as the von Neumann entropy of the density operator ρ.

10.8.2 Accessible information

Let us consider the scenario introduced at the beginning of this section, where Alice prepares an ensemble $\mathcal{E} \equiv \{p_X(x), \rho_x\}$ and Bob performs a POVM $\{\Lambda_y\}$. Suppose now that Bob is actually trying to retrieve as much information as possible about the random variable X. The quantity that governs how much information he can learn about random variable X if he possesses random variable Y is the mutual information $I(X;Y)$. But here, Bob can actually choose which measurement he would like to perform, and it would be good for

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

him to perform the measurement that maximizes his information about X. The resulting quantity is known as the accessible information $I_{\text{acc}}(\mathcal{E})$ of the ensemble $\mathcal{E}$ (because it is the information that Bob can access about random variable X):

$$I_{\text{acc}}(\mathcal{E}) \equiv \max_{\{\Lambda_y\}} I(X;Y), \tag{10.82}$$

where the marginal density $p_X(x)$ is that from the ensemble and the conditional density $p_{Y|X}(y|x)$ is given in (10.78). In the next chapter, we show how to obtain a natural bound on this quantity, called the *Holevo bound*. The bound arises from a quantum generalization of the data processing inequality.

10.8.3 Classical Mutual Information of a Bipartite State

A final quantity that we introduce is the classical mutual information $I_c(\rho^{AB})$ of a bipartite state ρ^{AB}. Suppose that Alice and Bob possess some bipartite state ρ^{AB} and would like to extract maximal classical correlation from it. That is, they each retrieve a random variable by performing respective local POVMs $\{\Lambda_x^A\}$ and $\{\Lambda_y^B\}$ on their halves of the bipartite state ρ^{AB}. These measurements produce respective random variables X and Y, and they would like X and Y to be as correlated as possible. A good measure of their resulting classical correlations obtainable from local quantum information processing is as follows:

$$I_c(\rho^{AB}) \equiv \max_{\{\Lambda_x^A\},\{\Lambda_y^B\}} I(X;Y), \tag{10.83}$$

where the joint distribution

$$p_{X,Y}(x,y) \equiv \text{Tr}\{(\Lambda_x^A \otimes \Lambda_y^B)\rho^{AB}\}. \tag{10.84}$$

Suppose that the state ρ^{AB} is classical, that is, it has the form

$$\rho^{AB} = \sum_{x,y} p_{X,Y}(x,y)|x\rangle\langle x|^A \otimes |y\rangle\langle y|^Y, \tag{10.85}$$

where the states $|x\rangle^A$ form an orthonormal basis and so do the states $|y\rangle^B$. Then, the optimal measurement in this case is for Alice to perform a von Neumann measurement in the basis $|x\rangle^A$ and inform Bob to perform a similar measurement in the basis $|y\rangle^B$. The amount of correlation they extract is then equal to $I(X;Y)$.

Exercise 10.8.1 Prove that it suffices to consider maximizing over rank-one POVMs when computing (10.83). (Hint: Consider refining the POVM $\{\Lambda_x\}$ as the rank-one POVM $\{|\phi_{x,z}\rangle\langle\phi_{x,z}|\}$, where we spectrally decompose Λ_x as $\sum_z |\phi_{x,z}\rangle\langle\phi_{x,z}|$, and then exploit the data processing inequality.)

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

10.9 History and Further Reading

The book of Cover and Thomas is an excellent introduction to entropy and information theory (some of the material in this chapter is similar to material appearing in that book) [57]. MacKay's book is also a good introduction [189]. E. T. Jaynes was an advocate of the Principle of Maximum Entropy, proclaiming its utility in several sources [161, 162, 163]. A good exposition of Fano's inequality appears on Scholarpedia [92].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 11

Quantum Information and Entropy

In this chapter, we discuss several information measures that are important for quantifying the amount of information and correlations that are present in quantum systems. The first fundamental measure that we introduce is the von Neumman entropy. It is the quantum analog of the Shannon entropy, but it captures both classical and quantum uncertainty in a quantum state.[1] The von Neumann entropy gives meaning to a notion of the *information qubit*. This notion is different from that of the physical qubit, which is the description of a quantum state in an electron or a photon. The information qubit is the fundamental quantum informational unit of measure, determining how much quantum information is in a quantum system.

The beginning definitions here are analogous to the classical definitions of entropy, but we soon discover a radical departure from the intuitive classical notions from the previous chapter: the conditional quantum entropy can be negative for certain quantum states. In the classical world, this negativity simply does not occur, though it takes a special meaning in quantum information theory. Pure quantum states that are entangled have stronger-than-classical spatial correlations and are examples of states that have negative conditional entropy. The negative of the conditional quantum entropy is so important in quantum information theory that we even have a special name for it: the coherent information. We discover that the coherent information obeys a quantum data processing inequality, placing it on a firm footing as a particular informational measure of quantum correlations.

We then define several other quantum information measures, such as quantum mutual information, that bear similar definitions as in the classical world, but with Shannon entropies replaced with von Neumann entropies. This replacement may seem to make quantum entropy

[1] We should point out the irony in the historical development of classical and quantum entropy. The von Neumann entropy has seen much widespread use in modern quantum information theory, and perhaps this would make one think that von Neumann discovered this quantity much after Shannon. But in fact, the reverse is true. Von Neumann first discovered what is now known as the von Neumann entropy and applied it to questions in statistical physics. Much later, Shannon determined an information-theoretic formula and asked von Neumann what he should call it. Von Neumann told him to call it the entropy for two reasons: 1) it was a special case of the von Neumann entropy and 2) he would always have the advantage in a debate because von Neumann claimed that no one at the time really understood entropy.

somewhat trivial on the surface, but a simple calculation reveals that a maximally entangled state on two qubits registers *two bits* of quantum mutual information (recall that the largest the mutual information can be in the classical world is *one bit* for the case of two maximally correlated bits). We then discuss several information inequalities that play an important role in quantum information processing: the fundamental quantum information inequality, strong subadditivity, the quantum data processing inequality, and continuity of quantum entropy.

11.1 Quantum Entropy

We might expect a measure of the entropy of a quantum system to be vastly different from the classical measure of entropy from the previous chapter because a quantum system possesses not only classical uncertainty but also quantum uncertainty that arises from the uncertainty principle. But recall that the density operator captures both types of uncertainty and allows us to determine probabilities for the outcomes of any measurement on system A. Thus, a quantum measure of uncertainty should be a direct function of the density operator, just as the classical measure of uncertainty is a direct function of a probability density function. It turns out that this function has a strikingly similar form to the classical entropy, as we see below.

Definition 11.1.1 (Quantum Entropy)**.** *Suppose that Alice prepares some quantum system A in a state ρ^A. Then the entropy $H(A)$ of the state is as follows:*

$$H(A) \equiv -\text{Tr}\{\rho^A \log \rho^A\}. \tag{11.1}$$

The entropy of a quantum system is also known as *the von Neumann entropy* or *the quantum entropy* but we often simply refer to it as *the entropy*. We can denote it by $H(A)_\rho$ or $H(\rho)$ to show the explicit dependence on the density operator ρ^A. The von Neumann entropy has a special relation to the eigenvalues of the density operator, as the following exercise asks you to verify.

Exercise 11.1.1 Consider a density operator ρ^A with the following spectral decomposition:

$$\rho^A = \sum_x p_X(x)|x\rangle\langle x|^A. \tag{11.2}$$

Show that the entropy $H(A)$ is the same as the entropy $H(X)$ of a random variable X with probability distribution $p_X(x)$.

In our definition of quantum entropy, we use the same notation H as in the classical case to denote the entropy of a quantum system. It should be clear from context whether we are referring to the entropy of a quantum or classical system.

The quantum entropy admits an intuitive interpretation. Suppose that Alice generates a random quantum state $|\psi_y\rangle$ in her lab according to some probability density $p_Y(y)$ of a

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

random variable Y. Suppose further that Bob has not yet received the state from Alice and does not know which one she sent. The expected density operator from Bob's point of view is then

$$\sigma = \mathbb{E}_Y\{|\psi_Y\rangle\langle\psi_Y|\} = \sum_y p_Y(y)|\psi_y\rangle\langle\psi_y|. \tag{11.3}$$

The interpretation of the entropy $H(\sigma)$ is that it quantifies Bob's uncertainty about the state Alice sent—his expected information gain is $H(\sigma)$ qubits upon receiving and measuring the state that Alice sends. Schumacher's noiseless quantum coding theorem, described in Chapter 17, gives an alternative operational interpretation of the von Neumann entropy by proving that Alice needs to send Bob qubits at a rate $H(\sigma)$ in order for him to be able to decode a compressed quantum state.

The above interpretation of quantum entropy seems qualitatively similar to the interpretation of classical entropy. Though, there is a significant quantitative difference that illuminates the difference between Shannon entropy and von Neumann entropy. We consider an example. Suppose that Alice generates a sequence $|\psi_1\rangle|\psi_2\rangle\cdots|\psi_n\rangle$ of quantum states according to the following "BB84" ensemble:

$$\{\{1/4, |0\rangle\}, \{1/4, |1\rangle\}, \{1/4, |+\rangle\}, \{1/4, |-\rangle\}\}. \tag{11.4}$$

Suppose that her and Bob share a noiseless classical channel. If she employs Shannon's classical noiseless coding protocol, she should transmit classical data to Bob at a rate of two classical channel uses per source state $|\psi_i\rangle$ in order for him to reliably recover the classical data needed to reproduce the sequence of states that Alice transmitted (the Shannon entropy of the uniform distribution 1/4 is 2 bits).

Now let us consider computing the von Neumann entropy of the above ensemble. First, we determine the expected density operator of Alice's ensemble:

$$\frac{1}{4}(|0\rangle\langle 0| + |1\rangle\langle 1| + |+\rangle\langle +| + |-\rangle\langle -|) = \pi, \tag{11.5}$$

where π is the maximally mixed state. The von Neumann entropy of the above density operator is one qubit because the eigenvalues of π are both equal to 1/2. Suppose now that Alice and Bob share a noiseless quantum channel between them—this is a channel that can preserve quantum coherence without any interaction with an environment. Then Alice only needs to send qubits at a rate of one channel use per source symbol if she employs a protocol known as Schumacher compression (we discuss this protocol in detail in Chapter 17). Bob can then reliably decode the qubits that Alice sent. The protocol also causes only an asymptotically vanishing disturbance to the state. The above departure from classical information theory holds in general—Exercise 11.9.2 of this chapter asks you to prove that the Shannon entropy of any ensemble is never less than the von Neumann entropy of its expected density operator.

11.1.1 Mathematical Properties of Quantum Entropy

We now discuss several mathematical properties of the quantum entropy: positivity, its minimum value, its maximum value, its invariance under unitaries, and concavity. The first

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

three of these properties follow from the analogous properties in the classical world because the von Neumann entropy of a density operator is the Shannon entropy of its eigenvalues (see Exercise 11.1.1). We state them formally below:

Property 11.1.1 (Positivity) The von Neumann entropy $H(\rho)$ is non-negative for any density operator ρ:

$$H(\rho) \geq 0. \tag{11.6}$$

Proof. Positivity of quantum entropy follows from positivity of Shannon entropy. □

Property 11.1.2 (Minimum Value) The minimum value of the von Neumann entropy is zero, and it occurs when the density operator is a pure state.

Proof. The minimum value equivalently occurs when the eigenvalues of a density operator are distributed with all the mass on one value and zero on the others, so that the density operator is rank one and corresponds to a pure state. □

Why should the entropy of a pure quantum state vanish? It seems that there is quantum uncertainty inherent in the state itself and that a measure of quantum uncertainty should capture this fact. This last observation only makes sense if we do not know anything about the state that is prepared. But if we know exactly how it was prepared, we can perform a special quantum measurement to verify that the quantum state was prepared, and we do not learn anything from this measurement because the outcome of it is always certain. For example, suppose that Alice always prepares the state $|\phi\rangle$ and Bob knows that she does so. He can then perform a measurement of the following form $\{|\phi\rangle\langle\phi|, I - |\phi\rangle\langle\phi|\}$ to verify that she prepared this state. He always receives the first outcome from the measurement and never gains any information from it. Thus, it make sense to say that the entropy of a pure state vanishes.

Property 11.1.3 (Maximum Value) The maximum value of the von Neumann entropy is $\log D$ where D is the dimension of the system, and it occurs for the maximally mixed state.

Proof. The proof of the above property is the same as in the classical case. □

Property 11.1.4 (Concavity) The entropy is concave in the density operator:

$$H(\rho) \geq \sum_x p_X(x) H(\rho_x), \tag{11.7}$$

where $\rho \equiv \sum_x p_X(x)\rho_x$.

The physical interpretation of concavity is as before for classical entropy: entropy can never decrease under a mixing operation. This inequality is a fundamental property of the entropy, and we prove it after developing some important entropic tools (see Exercise 11.6.9).

Property 11.1.5 (Unitary Invariance) The entropy of a density operator is invariant under unitary operations on it:

$$H(\rho) = H\big(U\rho U^\dagger\big). \tag{11.8}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Proof. Unitary invariance of entropy follows by observing that the eigenvalues of a density operator are invariant under a unitary:

$$U\rho U^\dagger = U\sum_x p_X(x)|x\rangle\langle x|U^\dagger \tag{11.9}$$

$$= \sum_x p_X(x)|\phi_x\rangle\langle\phi_x|, \tag{11.10}$$

where $\{|\phi_x\rangle\}$ is some orthonormal basis such that $U|x\rangle = |\phi_x\rangle$. The above property follows because the entropy is a function of the eigenvalues of a density operator. □

A unitary operator is the quantum analog of a permutation in this context (consider Property 10.1.3 of the classical entropy).

Exercise 11.1.2 The purity of a density operator ρ^A is $\text{Tr}\left\{\left(\rho^A\right)^2\right\}$. Suppose $\rho^A = \text{Tr}_B\left\{\left(\Phi^+\right)^{AB}\right\}$. Prove that the purity is equal to the inverse of the dimension d in this case.

11.1.2 Alternate Characterization of the von Neumann Entropy

There is an interesting alternate characterization of the von Neumann entropy of a state ρ as the minimum Shannon entropy of a rank-one POVM performed on it (we discussed this briefly in Section 10.8.1). That is, we would like to show that

$$H(\rho) = \min_{\{\Lambda_y\}} -\sum_y \text{Tr}\{\Lambda_y\rho\}\log_2(\text{Tr}\{\Lambda_y\rho\}), \tag{11.11}$$

where the minimum is restricted to be over rank-one POVMs (those with $\Lambda_y = |\phi_y\rangle\langle\phi_y|$ for some vectors $|\phi_y\rangle$ such that $\text{Tr}\{|\phi_y\rangle\langle\phi_y|\} \leq 1$ and $\sum_y |\phi_y\rangle\langle\phi_y| = I$). In this sense, there is some optimal measurement to perform on ρ such that its entropy is equivalent to the von Neumann entropy, and this optimal measurement is the "right question to ask" (as we discussed early on in Section 1.2.2).

In order to prove the above result, we should first realize that a von Neumann measurement in the eigenbasis of ρ should achieve the minimum. That is, if $\rho = \sum_x p_X(x)|x\rangle\langle x|$, we should expect that the measurement $\{|x\rangle\langle x|\}$ achieves the minimum. In this case, the Shannon entropy of the measurement is equal to the Shannon entropy of $p_X(x)$, as discussed in Exercise 11.1.1. We now prove that any other rank-one POVM has a higher entropy than that given by this measurement. Consider that the distribution of the measurement outcomes for $\{|\phi_y\rangle\langle\phi_y|\}$ is equal to

$$\text{Tr}\{|\phi_y\rangle\langle\phi_y|\rho\} = \sum_x |\langle\phi_y|x\rangle|^2 p_X(x), \tag{11.12}$$

so that we can think of $|\langle\phi_y|x\rangle|^2$ as a conditional probability distribution. Introducing

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$f(p) \equiv -p \log_2 p$, which is a concave function, we can write the von Neumann entropy as

$$H(\rho) = \sum_x f(p_X(x)) \tag{11.13}$$

$$= \sum_x f(p_X(x)) + f(p_X(x_0)), \tag{11.14}$$

where x_0 is a symbol added to the alphabet of x such that $p_X(x_0) = 0$. Let us denote the enlarged alphabet with the symbols x' so that $H(\rho) = \sum_{x'} f(p_X(x'))$. We know that $\sum_y |\langle \phi_y | x \rangle|^2 = 1$ from the fact that the set $\{|\phi_y\rangle\langle\phi_y|\}$ forms a POVM and $|x\rangle$ is a normalized state. We also know that $\sum_x |\langle \phi_y | x \rangle|^2 \leq 1$ because $\text{Tr}\{|\phi_y\rangle\langle\phi_y|\} \leq 1$ for a rank-one POVM. Thinking of $|\langle \phi_y | x \rangle|^2$ as a distribution over x, we can add a symbol x_0 with probability $1 - \langle \phi_y | \phi_y \rangle$ so that it makes a normalized distribution. Let us call this distribution $p(x'|y)$. We then have that

$$H(\rho) = \sum_x f(p_X(x)) \tag{11.15}$$

$$= \sum_{x,y} |\langle \phi_y | x \rangle|^2 f(p_X(x)) \tag{11.16}$$

$$= \sum_{x',y} p(x'|y) f(p_X(x')) \tag{11.17}$$

$$= \sum_y \left(\sum_{x'} p(x'|y) f(p_X(x')) \right) \tag{11.18}$$

$$\leq \sum_y f\left(\sum_{x'} p(x'|y) p_X(x') \right) \tag{11.19}$$

$$= \sum_y f(\text{Tr}\{|\phi_y\rangle\langle\phi_y|\rho\}). \tag{11.20}$$

The third equality follows from our assumption that $p_X(x_0)$ for the added symbol x_0. The only inequality follows from concavity of f. The last expression is equivalent to the Shannon entropy of the POVM $\{|\phi_y\rangle\langle\phi_y|\}$ when performed on the state ρ.

11.2 Joint Quantum Entropy

The joint quantum entropy $H(AB)_\rho$ of the density operator ρ^{AB} for a bipartite system AB follows naturally from the definition of quantum entropy:

$$H(AB)_\rho \equiv -\text{Tr}\{\rho^{AB} \log \rho^{AB}\}. \tag{11.21}$$

We introduce a few of its properties in the below subsections.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

11.2.1 Marginal Entropies of a Pure Bipartite State

The five properties of quantum entropy in the previous section may give you the impression that the nature of quantum information is not too different from that of classical information. We proved all these properties for the classical case, and their proofs for the quantum case seem similar. The first three even resort to the proofs in the classical case!

Theorem 11.2.1 below is where we observe our first radical departure from the classical world. It states that the marginal entropies of a pure bipartite state are equal, while the entropy of the overall state remains zero. Recall that the joint entropy $H(X,Y)$ of two random variables X and Y is never less than one of the marginal entropies $H(X)$ or $H(Y)$:

$$H(X,Y) \geq H(X), \tag{11.22}$$

$$H(X,Y) \geq H(Y). \tag{11.23}$$

The above inequalities follow from the positivity of classical conditional entropy. But in the quantum world, these inequalities do not always have to hold, and the following theorem demonstrates that they do not hold for an arbitrary pure bipartite quantum state with Schmidt rank greater than one (see Theorem 3.6.1 for a definition of Schmidt rank). The fact that the joint quantum entropy can be less than the marginal quantum entropy contains in it one of the most fundamental differences between classical and quantum information.

Theorem 11.2.1. *The marginal entropies $H(A)_\phi$ and $H(B)_\phi$ of a pure bipartite state $|\phi\rangle^{AB}$ are equal:*

$$H(A)_\phi = H(B)_\phi, \tag{11.24}$$

while the joint entropy $H(AB)_\phi$ vanishes:

$$H(AB)_\phi = 0. \tag{11.25}$$

Proof. The crucial ingredient for the proof of this theorem is the Schmidt decomposition (Theorem 3.6.1). Recall that any bipartite state $|\phi\rangle^{AB}$ admits a Schmidt decomposition of the following form:

$$|\phi\rangle^{AB} = \sum_i \sqrt{\lambda_i} |i\rangle^A |i\rangle^B, \tag{11.26}$$

where $|i\rangle^A$ is some orthonormal set of vectors on system A and $|i\rangle^B$ is some orthonormal set on system B. Recall that the Schmidt rank is equal to the number of non-zero coefficients λ_i. Then the respective marginal states ρ^A and ρ^B on systems A and B are as follows:

$$\rho^A = \sum_i \lambda_i |i\rangle\langle i|^A, \tag{11.27}$$

$$\rho^B = \sum_i \lambda_i |i\rangle\langle i|^B. \tag{11.28}$$

Thus, the marginal states admit a spectral decomposition with the same eigenvalues. The theorem follows because the von Neumann entropy depends only on the eigenvalues of a given spectral decomposition. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The theorem applies not only to two systems A and B, but it also applies to any number of systems if we make a bipartite cut of the systems. For example, if the state is $|\phi\rangle^{ABCDE}$, then the following equalities (and others from different combinations) hold by applying Theorem 11.2.1 and Remark 3.6.1:

$$H(A)_\phi = H(BCDE)_\phi \tag{11.29}$$

$$H(AB)_\phi = H(CDE)_\phi \tag{11.30}$$

$$H(ABC)_\phi = H(DE)_\phi \tag{11.31}$$

$$H(ABCD)_\phi = H(E)_\phi \tag{11.32}$$

The closest analogy in the classical world to the above property is when we copy a random variable X. That is, suppose that X has a distribution $p_X(x)$ and $\hat{X}$ is some copy of it so that the distribution of the joint random variable $X\hat{X}$ is $p_X(x)\delta_{x,\hat{x}}$. Then the marginal entropies $H(X)$ and $H(\hat{X})$ are both equal. But observe that the joint entropy $H(X\hat{X})$ is also equal to $H(X)$ and this is where the analogy breaks down.

11.2.2 Additivity

The quantum entropy is additive for tensor product states:

$$H(\rho \otimes \sigma) = H(\rho) + H(\sigma). \tag{11.33}$$

One can verify this property simply by diagonalizing both density operators and resorting to the additivity of the joint Shannon entropies of the eigenvalues.

Additivity is a property that we would like to hold for any measure of information. For example, suppose that Alice generates a large sequence $|\psi_{x_1}\rangle|\psi_{x_2}\rangle \cdots |\psi_{x_n}\rangle$ of quantum states according to the ensemble $\{p_X(x), |\psi_x\rangle\}$. She may be aware of the classical indices $x_1 x_2 \cdots x_n$, but a third party to whom she sends the quantum sequence may not be aware of these values. The description of the state to this third party is then as follows:

$$\rho \otimes \cdots \otimes \rho, \tag{11.34}$$

where $\rho \equiv \mathbb{E}_X\{|\psi_X\rangle\langle\psi_X|\}$, and the quantum entropy of this n-fold tensor product state is

$$H(\rho \otimes \cdots \otimes \rho) = nH(\rho), \tag{11.35}$$

by applying (11.33) inductively.

11.2.3 Joint Quantum Entropy of a Classical-Quantum State

Recall that a classical-quantum state is a bipartite state in which a classical system and a quantum system are classically correlated. An example of such a state is as follows:

$$\rho^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^B. \tag{11.36}$$

The joint quantum entropy of this state takes on a special form that appears similar to entropies in the classical world.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Theorem 11.2.2. *The joint entropy $H(XB)_\rho$ of a classical-quantum state is as follows:*

$$H(XB)_\rho = H(X) + \sum_x p_X(x) H(\rho_x), \tag{11.37}$$

where $H(X)$ is the entropy of a random variable with distribution $p_X(x)$.

Proof. First, suppose that the conditional density operators ρ_x^B have the following spectral decomposition:

$$\rho_x^B = \sum_y p_{Y|X}(y|x) |y_x\rangle\langle y_x|^B, \tag{11.38}$$

where we write the eigenstates $|y_x\rangle$ with a subscript x to indicate that the basis $\{|y_x\rangle\}$ may be different for each value of x. Then

$$\begin{aligned}
&H(XB)_\rho \\
&= -\text{Tr}\{\rho^{XB} \log \rho^{XB}\} && (11.39)\\
&= -\text{Tr}\left\{ \begin{array}{l} \sum_{x,y} p_X(x) p_{Y|X}(y|x) |x\rangle\langle x|^X \otimes |y_x\rangle\langle y_x|^B \times \\ \quad \left(\log \sum_{x',y'} p_X(x') p_{Y|X}(y'|x') |x'\rangle\langle x'|^X \otimes |y'_{x'}\rangle\langle y'_{x'}|^B\right) \end{array} \right\} && (11.40)\\
&= -\sum_{x,y,x',y'} p_X(x) p_{Y|X}(y|x) \log\big(p_X(x') p_{Y|X}(y'|x')\big) \times \\
&\quad \text{Tr}\Big\{ |x\rangle\langle x|x'\rangle\langle x'|^X \otimes |y_x\rangle\langle y_x|y'_{x'}\rangle\langle y'_{x'}|^B \Big\} && (11.41)
\end{aligned}$$

The first equality follows by definition. The second equality follows by expanding ρ^{XB} with (11.36) and (11.38). The third equality follows because $f(A) = f(\sum_i a_i |i\rangle\langle i|) = \sum_i f(a_i) |i\rangle\langle i|$ where $\sum_i a_i |i\rangle\langle i|$ is a spectral decomposition of A. Continuing,

$$\begin{aligned}
&= -\sum_{x,y,y'} p_X(x) p_{Y|X}(y|x) \log\big(p_X(x) p_{Y|X}(y'|x)\big) \text{Tr}\{ |y_x\rangle\langle y_x|y'_x\rangle\langle y'_x|^B \} && (11.42)\\
&= -\sum_{x,y} p_X(x) p_{Y|X}(y|x) \log\big(p_X(x) p_{Y|X}(y|x)\big) \text{Tr}\Big\{ |y_x\rangle\langle y_x|^B \Big\} && (11.43)\\
&= -\sum_{x,y} p_X(x) p_{Y|X}(y|x) \log\big(p_X(x) p_{Y|X}(y|x)\big) && (11.44)\\
&= -\sum_{x,y} p_X(x) \log(p_X(x)) - \sum_x p_X(x) \sum_y p_{Y|X}(y|x) \log\big(p_{Y|X}(y|x)\big) && (11.45)\\
&= H(X) + \sum_x p_X(x) H(\rho_x). && (11.46)
\end{aligned}$$

The first equality follows from linearity of trace and partially evaluating it. The second equality follows because the eigenstates $\{|y_x\rangle\}$ form an orthonormal basis for the same x. The third equality follows because $\text{Tr}\Big\{ |y_x\rangle\langle y_x|^B \Big\} = 1$. The fourth equality follows because

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

the logarithm of the products is the sum of the logarithms. The final equality follows from the definition of entropy and because $-\sum_y p_{Y|X}(y|x)\log\big(p_{Y|X}(y|x)\big)$ is the quantum entropy of the density operator ρ_x. □

As we stated earlier, the joint quantum entropy of a classical-quantum state takes on a special form that is analogous to the classical joint entropy. Inspection of the third to last line above reveals this similarity because it looks exactly the same as the formula in (10.23). We explore this connection in further detail in Section 11.4.1.

11.3 Potential yet Unsatisfactory Definitions of Conditional Quantum Entropy

The conditional quantum entropy may perhaps seem a bit difficult to define at first because there is no formal notion of conditional probability in the quantum theory. Though, there are two senses which are perhaps closest to the notion of conditional probability, but both of them do not lead to satisfactory definitions of conditional quantum entropy. Nevertheless, it is instructive for us to explore both of these notions for a bit. The first arises in the noisy quantum theory, and the second arises in the purified quantum theory.

We develop the first notion. Consider an arbitrary bipartite state ρ^{AB}. Suppose that Alice performs a complete von Neumann measurement $\Pi \equiv \{|x\rangle\langle x|\}$ of her system in the basis $\{|x\rangle\}$. This procedure leads to an ensemble $\{p_X(x), |x\rangle\langle x|^A \otimes \rho_x\}$, where

$$\rho_x \equiv \frac{1}{p_X(x)}\text{Tr}_A\Big\{\Big(|x\rangle\langle x|^A \otimes I^B\Big)\rho^{AB}\Big(|x\rangle\langle x|^A \otimes I^B\Big)\Big\}, \tag{11.47}$$

$$p_X(x) \equiv \text{Tr}\Big\{\Big(|x\rangle\langle x|^A \otimes I^B\Big)\rho^{AB}\Big\}. \tag{11.48}$$

One could then think of the density operators ρ_x as being conditional on the outcome of the measurement, and these density operators describe the state of Bob given knowledge of the outcome of the measurement.

We could potentially define a conditional entropy as follows:

$$H(B|A)_\Pi \equiv \sum_x p_X(x)H(\rho_x), \tag{11.49}$$

in analogy with the definition of the classical entropy in (10.18). This approach might seem to lead to a useful definition of conditional quantum entropy, but the problem with it is that the entropy depends on the measurement chosen (the notation $H(B|A)_\Pi$ explicitly indicates this dependence). This problem does not occur in the classical world because the probabilities for the outcomes of measurements do not themselves depend on the measurement selected, unless we apply some coarse graining to the outcomes. Though, this dependence on measurement is a fundamental aspect of the quantum theory.

We could then attempt to remove the dependence of the above definition on a particular measurement Π by defining the conditional quantum entropy to be the minimization of

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$H(B|A)_{\Pi}$ over all possible measurements. The intuition here is perhaps that entropy should be the minimal amount of conditional uncertainty in a system after employing the best possible measurement on the other. Though, the removal of one problem leads to another! This optimized conditional entropy is now difficult to compute as the system grows larger, whereas in the classical world, the computation of conditional entropy is simple if one knows the conditional probabilities. The above idea is useful, but we leave it for now because there is a simpler definition of conditional quantum entropy that plays a fundamental role in quantum information theory.

The second notion of conditional probability is actually similar to the above notion, though we present it in the purified viewpoint. Consider a tripartite state $|\psi\rangle^{ABC}$ and a bipartite cut $A|BC$ of the systems A, B, and C. Theorem 3.6.1 states that every bipartite state admits a Schmidt decomposition, and the state $|\psi\rangle^{ABC}$ is no exception. Thus, we can write a Schmidt decomposition for it as follows:

$$|\psi\rangle^{ABC} = \sum_x \sqrt{p_X(x)}|x\rangle^A|\phi_x\rangle^{BC}, \tag{11.50}$$

where $p_X(x)$ is some probability density, $\{|x\rangle\}$ is an orthonormal basis for the system A, and $\{|\phi_x\rangle\}$ is an orthonormal basis for the systems BC. Each state $|\phi_x\rangle^{BC}$ is a pure bipartite state, so we can again apply a Schmidt decomposition to each of these states:

$$|\phi_x\rangle^{BC} = \sum_y \sqrt{p_{Y|X}(y|x)}|y_x\rangle^B|y_x\rangle^C, \tag{11.51}$$

where $p_{Y|X}(y|x)$ is some conditional probability distribution depending on the value of x, and $\{|y_x\rangle^B\}$ and $\{|y_x\rangle^C\}$ are both orthonormal bases with dependence on the value x. Thus, the overall state has the following form:

$$|\psi\rangle^{ABC} = \sum_{x,y} \sqrt{p_{Y|X}(y|x)p_X(x)}|x\rangle^A|y_x\rangle^B|y_x\rangle^C. \tag{11.52}$$

Suppose that Alice performs a von Neumann measurement in the basis $\{|x\rangle\langle x|^A\}$. The state on Bob and Charlie's systems is then $|\psi_x\rangle^{BC}$, and each system on B or C has a marginal entropy of $H(\sigma_x)$ where $\sigma_x \equiv \sum_y p_{Y|X}(y|x)|y_x\rangle\langle y_x|$. We could potentially define the conditional quantum entropy as

$$\sum_x p_X(x)H(\sigma_x). \tag{11.53}$$

This quantity does not depend on a measurement as before because we simply choose the measurement from the Schmidt decomposition. But there are many problems with the above notion of conditional quantum entropy: it is defined only for pure quantum states, it is not clear how to apply it to a bipartite quantum system, and the conditional entropy of Bob's system given Alice's and that of Charlie's given Alice's is the same (which is perhaps the strangest of all!). Thus this notion of conditional probability is not useful for a definition of conditional quantum entropy.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

11.4 Conditional Quantum Entropy

The definition of conditional quantum entropy that has been most useful in quantum information theory is the following simple one, inspired from the relation between joint entropy and marginal entropy in Exercise 10.3.1.

Definition 11.4.1 (Conditional Quantum Entropy)**.** *The conditional quantum entropy $H(A|B)_\rho$ of a bipartite quantum state ρ^{AB} is the difference of the joint quantum entropy $H(AB)_\rho$ and the marginal $H(B)_\rho$:*

$$H(A|B)_\rho = H(AB)_\rho - H(B)_\rho. \tag{11.54}$$

The above definition is the most natural one, both because it is straightforward to compute for any bipartite state and because it obeys many relations that the classical conditional entropy obeys (such as chaining rules and conditioning reduces entropy). We explore many of these relations in the forthcoming sections. For now, we state "conditioning cannot increase entropy" as the following theorem and tackle its proof later on after developing a few more tools.

Theorem 11.4.1 (Conditioning does not increase entropy)**.** *Consider a bipartite quantum state ρ^{AB}. Then the following inequality applies to the marginal entropy $H(A)_\rho$ and the conditional quantum entropy $H(A|B)_\rho$:*

$$H(A)_\rho \geq H(A|B)_\rho. \tag{11.55}$$

We can interpret the above inequality as stating that conditioning cannot increase entropy, even if the conditioning system is quantum.

11.4.1 Conditional Quantum Entropy for Classical-Quantum States

A classical-quantum state is an example of a state where conditional quantum entropy behaves as in the classical world. Suppose that two parties share a classical-quantum state ρ^{XB} of the form in (11.36). The system X is classical and the system B is quantum, and the correlations between these systems are entirely classical, determined by the probability distribution $p_X(x)$. Let us calculate the conditional quantum entropy $H(B|X)_\rho$ for this state:

$$\begin{aligned} H(B|X)_\rho &= H(XB)_\rho - H(X)_\rho && (11.56)\\ &= H(X) + \sum_x p_X(x) H(\rho_x) - H(X) && (11.57)\\ &= \sum_x p_X(x) H(\rho_x). && (11.58) \end{aligned}$$

The first equality follows from Definition 11.4.1. The second equality follows from Theorem 11.2.2, and the final equality results from algebra.

The above form for conditional entropy is completely analogous with the classical formula in (10.18) and holds whenever the conditioning system is classical.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

11.4.2 Negative Conditional Quantum Entropy

One of the properties of the conditional quantum entropy in Definition 11.4.1 that seems counterintuitive at first sight is that it can be negative. This negativity holds for an ebit $|\Phi^+\rangle^{AB}$ shared between Alice and Bob. The marginal state on Bob's system is the maximally mixed state π^B. Thus, the marginal entropy $H(B)$ is equal to one, but the joint entropy vanishes, and so the conditional quantum entropy $H(A|B) = -1$.

What do we make of this result? Well, this is one of the fundamental differences between the classical world and the quantum world, and perhaps is the very essence of the departure from an informational standpoint. The informational statement is that we can sometimes be more certain about the joint state of a quantum system than we can be about any one of its individual parts, and this is the reason that conditional quantum entropy can be negative. This is in fact the same observation that Schrödinger made concerning entangled states [215]:

> "When two systems, of which we know the states by their respective representatives, enter into temporary physical interaction due to known forces between them, and when after a time of mutual influence the systems separate again, then they can no longer be described in the same way as before, viz. by endowing each of them with a representative of its own. I would not call that one but rather the characteristic trait of quantum mechanics, the one that enforces its entire departure from classical lines of thought. By the interaction the two representatives [the quantum states] have become entangled. Another way of expressing the peculiar situation is: the best possible knowledge of a whole does not necessarily include the best possible knowledge of all its parts, even though they may be entirely separate and therefore virtually capable of being 'best possibly known,' i.e., of possessing, each of them, a representative of its own. The lack of knowledge is by no means due to the interaction being insufficiently known — at least not in the way that it could possibly be known more completely — it is due to the interaction itself."

These explanations might aid somewhat in understanding a negative conditional entropy, but the ultimate test for whether we truly understand an information measure is if it is the answer to some operational task. The task where we can interpret the conditional quantum entropy is known as *state merging*. Suppose that Alice and Bob share n copies of a bipartite state ρ^{AB} where n is a large number and A and B are qubit systems. We also allow them free access to a classical side channel, but we count the number of times that they use a noiseless qubit channel. Alice would like to send Bob qubits over a noiseless qubit channel so that he receives her share of the state ρ^{AB}, i.e., so that he possesses all of the A shares. The naive approach would be for Alice simply to send her shares of the state over the noiseless qubit channels, i.e., she would use the channel n times to send all n shares. But the state merging protocol allows her to do much better, depending on the state ρ^{AB}. If the state ρ^{AB} has positive conditional quantum entropy, she needs to use the noiseless qubit channel only $nH(A|B)$ times (we will prove later that $H(A|B) \leq 1$ for any bipartite state on qubit systems). Though, if the conditional quantum entropy is negative, she does not

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

need to use the noiseless qubit channel at all, and at the end of the protocol, Alice and Bob share $nH(A|B)$ noiseless ebits! They can then use these ebits for future communication purposes, such as a teleportation or super-dense coding protocol (see Chapter 6). Thus, a negative conditional quantum entropy implies that Alice and Bob gain the potential for future quantum communication, making clear in an operational sense what a negative conditional quantum entropy means.[2] (We will cover this protocol in Chapter 21).

Exercise 11.4.1 Show that $H(A|B)_\rho = H(A|BC)_\sigma$ if $\sigma^{ABC} = \rho^{AB} \otimes \tau^C$.

11.5 Coherent Information

Negativity of the conditional quantum entropy is so important in quantum information theory that we even have an information quantity and a special notation to denote the negative of the conditional quantum entropy:

Definition 11.5.1 (Coherent Information)**.** *The coherent information $I(A\rangle B)_\rho$ of a bipartite state ρ^{AB} is as follows:*

$$I(A\rangle B)_\rho \equiv H(B)_\rho - H(AB)_\rho. \tag{11.59}$$

You should immediately notice that this quantity is the negative of the conditional quantum entropy in Definition 11.4.1, but it is perhaps more useful to think of the coherent information not merely as the negative of the conditional quantum entropy, but as an information quantity in its own right. This is why we employ a separate notation for it. The "I" is present because the coherent information is an information quantity that measures quantum correlations, much like the mutual information does in the classical case. For example, we have already seen that the coherent information of an ebit is equal to one. Thus, it is measuring the extent to which we know less about part of a system than we do about its whole. Perhaps surprisingly, the coherent information obeys a quantum data processing inequality (discussed in Section 11.9.3), which gives further support for it having an "I" present in its notation. The Dirac symbol "$\rangle$" is present to indicate that this quantity is a quantum information quantity, having meaning really only in the quantum world. The choice of "$\rangle$" over "$\langle$" also indicates a directionality from Alice to Bob, and this notation will make more sense when we begin to discuss the coherent information of a quantum channel in Chapter 12.

Exercise 11.5.1 Calculate the coherent information $I(A\rangle B)_\Phi$ of the maximally entangled state

$$|\Phi\rangle^{AB} \equiv \frac{1}{\sqrt{D}} \sum_{i=1}^{D} |i\rangle^A |i\rangle^B. \tag{11.60}$$

[2] After Horodecki, Oppenheim, and Winter published the state merging protocol [148], the *Bristol Evening Post* featured a story about Andreas Winter with the amusing title "Scientist Knows Less Than Nothing," as a reference to the potential negativity of conditional quantum entropy. Of course, such a title may seem a bit non-sensical to the layman, but it does grasp the idea that we can know less about a part of a quantum system than we do about its whole.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Calculate the coherent information $I(A\rangle B)_{\overline{\Phi}}$ of the maximally correlated state

$$\overline{\Phi}^{AB} \equiv \frac{1}{D}\sum_{i=1}^{D} |i\rangle\langle i|^A \otimes |i\rangle\langle i|^B. \tag{11.61}$$

Exercise 11.5.2 Consider a bipartite state ρ^{AB}. Consider a purification of this state to some environment system E. Show that

$$I(A\rangle B)_\rho = H(B)_\rho - H(E)_\rho. \tag{11.62}$$

Thus, there is a sense in which the coherent information measures the difference in the uncertainty of Bob and the uncertainty of the environment.

Exercise 11.5.3 Show that $I(A\rangle B) = H(A|E)$ for the purification in the above exercise.

The coherent information can be both negative and positive depending on the bipartite state on which we evaluate it, but it cannot be arbitrarily large. The following theorem places a useful bound on its absolute value.

Theorem 11.5.1. *Suppose that Alice and Bob share a bipartite state ρ^{AB}. The following bound applies to the absolute value of the conditional entropy $H(A|B)$:*

$$|H(A|B)| \leq \log d_A, \tag{11.63}$$

where d_A is the dimension of Alice's system.

Proof. We first prove the inequality $H(A|B) \leq \log d_A$ in two steps:

$$\begin{aligned} H(A|B) &\leq H(A) && (11.64)\\ &\leq \log d_A. && (11.65)\end{aligned}$$

The first inequality follows because conditioning reduces entropy (Theorem 11.4.1), and the second inequality follows because the maximum value of the entropy $H(A)$ is $\log d_A$. We now prove the inequality $H(A|B) \geq -\log d_A$. Consider a purification $|\psi\rangle^{EAB}$ of the state ρ^{AB}. We then have that

$$\begin{aligned} H(A|B) &= -H(A|E) && (11.66)\\ &\geq -H(A) && (11.67)\\ &\geq -\log d_A. && (11.68)\end{aligned}$$

The first equality follows from Exercise 11.5.3. The first and second inequalities follow by the same reasons as the inequalities in the previous paragraph. □

Exercise 11.5.4 (Conditional Coherent Information) Consider a tripartite state ρ^{ABC}. Show that

$$I(A\rangle BC)_\rho = I(A\rangle B|C)_\rho, \tag{11.69}$$

where $I(A\rangle B|C)_\rho \equiv H(B|C)_\rho - H(AB|C)_\rho$ is the conditional coherent information.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 11.5.5 (Conditional Coherent Information of a Classical-Quantum State) Suppose we have a classical-quantum state σ^{XAB} where

$$\sigma^{XAB} = \sum_x p_X(x)|x\rangle\langle x| \otimes \sigma_x^{AB}. \tag{11.70}$$

Show that

$$I(A\rangle BX)_{\sigma^{XAB}} = \sum_x p_X(x) I(A\rangle B)_{\sigma_x^{AB}} \tag{11.71}$$

11.6 Quantum Mutual Information

The standard informational measure of correlations in the classical world is the mutual information, and such a quantity plays a prominent role in measuring classical and quantum correlations in the quantum world as well.

Definition 11.6.1 (Quantum Mutual Information)**.** *The quantum mutual information of a bipartite state ρ^{AB} is as follows:*

$$I(A;B)_\rho \equiv H(A)_\rho + H(B)_\rho - H(AB)_\rho. \tag{11.72}$$

The following relations hold for quantum mutual information, in analogy with the classical case:

$$\begin{aligned} I(A;B)_\rho &= H(A)_\rho - H(A|B)_\rho & (11.73)\\ &= H(B)_\rho - H(B|A)_\rho. & (11.74) \end{aligned}$$

These immediately lead to the following relations between quantum mutual information and the coherent information:

$$\begin{aligned} I(A;B)_\rho &= H(A)_\rho + I(A\rangle B)_\rho & (11.75)\\ &= H(B)_\rho + I(B\rangle A)_\rho & (11.76) \end{aligned}$$

The below theorem gives a fundamental lower bound on the quantum mutual information—we merely state it for now and give a full proof later.

Theorem 11.6.1 (Positivity of Quantum Mutual Information)**.** *The quantum mutual information $I(A;B)_\rho$ of any bipartite quantum state ρ^{AB} is positive:*

$$I(A;B)_\rho \geq 0. \tag{11.77}$$

Exercise 11.6.1 (Proof that conditioning does not increase entropy) Show that positivity of quantum mutual information implies that conditioning does not increase entropy (Theorem 11.4.1).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 11.6.2 Calculate the quantum mutual information $I(A;B)_{\Phi}$ of the maximally entangled state Φ^{AB}. Calculate the quantum mutual information $I(A;B)_{\overline{\Phi}}$ of the maximally correlated state $\overline{\Phi}^{AB}$.

Exercise 11.6.3 (Bound on Quantum Mutual Information) Prove that the following bound applies to the quantum mutual information:

$$I(A;B) \leq 2\min\{\log d_A, \log d_B\}, \tag{11.78}$$

where d_A is the dimension of system A and d_B is the dimension of system B.

Exercise 11.6.4 Consider a pure state $|\psi\rangle^{RA}$. Suppose that an isometry $\mathcal{U}^{A\to BE}$ acts on the A system to produce the state $|\phi\rangle^{RBE}$. Show that

$$I(R;B)_{\phi} + I(R;E)_{\phi} = I(R;A)_{\psi}. \tag{11.79}$$

Exercise 11.6.5 Consider a tripartite state $|\psi\rangle^{SRA}$. Suppose that an isometry $\mathcal{U}^{A\to BE}$ acts on the A system to produce the state $|\phi\rangle^{SRBE}$. Show that

$$I(R;A)_{\psi} + I(R;S)_{\psi} = I(R;B)_{\phi} + I(R;SE)_{\phi} \tag{11.80}$$

Exercise 11.6.6 (Entropy, Coherent Information, and Quantum Mutual Information) Consider a pure state $|\phi\rangle^{ABE}$ on systems ABE. Using the Schmidt decomposition with respect to the bipartite cut $A \mid BE$, we can write $|\phi\rangle^{ABE}$ as follows:

$$|\phi\rangle^{ABE} = \sum_x \sqrt{p_X(x)}|x\rangle^A \otimes |\phi_x\rangle^{BE}, \tag{11.81}$$

for some orthonormal states $\{|x\rangle^A\}_{x\in\mathcal{X}}$ on system A and some orthonormal states $\{|\phi_x\rangle^{BE}\}$ on the joint system BE. Prove the following relations:

$$I(A\rangle B)_{\phi} = \frac{1}{2}I(A;B)_{\phi} - \frac{1}{2}I(A;E)_{\phi}, \tag{11.82}$$

$$H(A)_{\phi} = \frac{1}{2}I(A;B)_{\phi} + \frac{1}{2}I(A;E)_{\phi}. \tag{11.83}$$

Exercise 11.6.7 (Coherent Information and Private Information) We obtain a decohered version $\overline{\phi}^{ABE}$ of the state in Exercise 11.6.6 by measuring the A system in the basis $\{|x\rangle^A\}_{x\in\mathcal{X}}$. Let us now denote the A system as the X system because it becomes a classical system after the measurement:

$$\overline{\phi}^{XBE} = \sum_x p_X(x)|x\rangle\langle x|^X \otimes \phi_x^{BE}. \tag{11.84}$$

Prove the following relation:

$$I(A\rangle B)_{\phi} = I(X;B)_{\overline{\phi}} - I(X;E)_{\overline{\phi}}. \tag{11.85}$$

The quantity on the RHS is known as the private information, because there is a sense in which it quantifies the classical information in X that is accessible to Bob while being private from Eve.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

11.6.1 Holevo information

Suppose that Alice prepares some classical ensemble $\mathcal{E} \equiv \{p_X(x), \rho_x^B\}$ and then hands this ensemble to Bob without telling him the classical index x. The expected density operator of this ensemble is

$$\rho^B \equiv \mathbb{E}_X\{\rho_X^B\} = \sum_x p_X(x)\rho_x^B, \tag{11.86}$$

and this density operator ρ^B characterizes the state from Bob's perspective because he does not have knowledge of the classical index x. His task is to determine the classical index x by performing some measurement on his system B. Recall from Section 10.8.2 that the accessible information quantifies Bob's information gain after performing some optimal measurement $\{\Lambda_y\}$ on his system B:

$$I_{\text{acc}}(\mathcal{E}) = \max_{\{\Lambda_y\}} I(X;Y), \tag{11.87}$$

where Y is a random variable corresponding to the outcome of the measurement.

What is the accessible information of the ensemble? In general, this quantity is difficult to compute, but another quantity, called the Holevo information, provides a useful upper bound. The Holevo information $\chi(\mathcal{E})$ of the ensemble is

$$\chi(\mathcal{E}) \equiv H(\rho^B) - \sum_x p_X(x) H(\rho_x^B). \tag{11.88}$$

Exercise 11.9.1 asks you to prove this upper bound after we develop the quantum data processing inequality for quantum mutual information. The Holevo information characterizes the correlations between the classical variable X and the quantum system B.

Exercise 11.6.8 (Quantum Mutual Information of Classical-Quantum States) Consider the following classical-quantum state representing the ensemble $\mathcal{E}$:

$$\sigma^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^B. \tag{11.89}$$

Show that the Holevo information $\chi(\mathcal{E})$ is equivalent to the mutual information $I(X;B)_\sigma$:

$$\chi(\mathcal{E}) = I(X;B)_\sigma. \tag{11.90}$$

In this sense, the quantum mutual information of a classical-quantum state is most similar to the classical mutual information of Shannon.

Exercise 11.6.9 (Concavity of Quantum Entropy) Prove the concavity of entropy (Property 11.1.4) using Theorem 11.6.1 and the result of Exercise 11.6.8.

Exercise 11.6.10 Prove that the following bound applies to the Holevo information:

$$I(X;B)_\sigma \leq \log d_X, \tag{11.91}$$

where d_X is the dimension of the random variable X and the quantum mutual information is with respect to the classical-quantum state in Exercise 11.6.8.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

11.7 Conditional Quantum Mutual Information

We define the conditional quantum mutual information $I(A;B|C)_\rho$ of any tripartite state ρ^{ABC} similarly to how we did in the classical case:

$$I(A;B|C)_\rho \equiv H(A|C)_\rho + H(B|C)_\rho - H(AB|C)_\rho. \tag{11.92}$$

One can exploit the above definition and the definition of quantum mutual information to prove a chain rule for quantum mutual information.

Property 11.7.1 (Chain Rule for Quantum Mutual Information) The quantum mutual information obeys a chain rule:

$$I(AB;C) = I(B;C|A) + I(A;C). \tag{11.93}$$

Exercise 11.7.1 Use the chain rule for quantum mutual information to prove the following relationship:

$$I(A;BC) = I(AC;B) + I(A;C) - I(B;C). \tag{11.94}$$

11.7.1 Positivity of Conditional Quantum Mutual Information

In the classical world, positivity of conditional mutual information follows trivially from positivity of mutual information (recall Theorem 10.6.1). The proof of positivity of conditional quantum mutual information is far from trivial in the quantum world, unless the conditioning system is classical (see Exercise 11.7.2). It is a wonderful thing that positivity of this quantity holds because so much of quantum information theory rests upon this theorem's shoulders (in fact, we could say that this inequality is one of the "bedrocks" of quantum information theory). The list of its corollaries includes the quantum data processing inequality, the answers to some additivity questions in quantum Shannon theory, the Holevo bound, and others. The proof of Theorem 11.7.1 follows directly from monotonicity of quantum relative entropy (Theorem 11.9.1), which we prove partially in the proof of Theorem 11.9.1 and fully in Appendix B of this book.

Theorem 11.7.1 (Positivity of Conditional Quantum Mutual Information)**.** *Suppose we have a quantum state on three systems A, B, and C. Then the conditional quantum mutual information is positive:*

$$I(A;B|C) \geq 0. \tag{11.95}$$

This condition is equivalent to the strong subadditivity inequality in Exercise 11.7.6, so we might also refer to the above inequality as strong subadditivity.

Exercise 11.7.2 (Conditional Quantum Mutual Information of Classical-Quantum States) Consider a classical-quantum state σ^{XAB} of the form in (11.70). Prove the following relation:

$$I(A;B|X)_\sigma = \sum_x p_X(x) I(A;B)_{\sigma_x}. \tag{11.96}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Conclude that positivity of conditional quantum mutual information is trivial in this special case where the conditioning system is classical, simply by exploiting positivity of quantum mutual information (Theorem 11.6.1).

Exercise 11.7.3 (Conditioning Does Not Increase Entropy) Consider a tripartite state ρ^{ABC}. Show that Theorem 11.7.1 implies the following stronger form of Theorem 11.4.1:

$$H(A|B)_\rho \geq H(A|BC)_\rho. \tag{11.97}$$

Exercise 11.7.4 (Concavity of Conditional Quantum Entropy) Show that strong subadditivity implies that conditional entropy is concave. That is, prove that

$$\sum_x p_X(x) H(A|B)_{\rho_x} \leq H(A|B)_\rho, \tag{11.98}$$

where $\rho^{AB} \equiv \sum_x p_X(x)\rho_x^{AB}$.

Exercise 11.7.5 (Convexity of Coherent Information) Prove that coherent information is convex:

$$\sum_x p_X(x) I(A\rangle B)_{\rho_x} \geq I(A\rangle B)_\rho, \tag{11.99}$$

by exploiting the result of the above exercise.

Exercise 11.7.6 (Strong Subadditivity) Theorem 11.7.1 also goes by the name of "strong subadditivity" because it is an example of a function ϕ that is strongly subadditive:

$$\phi(E) + \phi(F) \geq \phi(E \cap F) + \phi(E \cup F). \tag{11.100}$$

Show that positivity of quantum conditional mutual information implies the following strong subadditivity relation:

$$H(AB) + H(BC) \geq H(B) + H(ABC), \tag{11.101}$$

where we think of ϕ in (11.100) as the entropy function H, the argument E in (11.100) as AB, and the argument F in (11.100) as BC.

11.8 Quantum Relative Entropy

The quantum relative entropy $D(\rho \,||\, \sigma)$ between two states ρ and σ is as follows:

$$D(\rho \,||\, \sigma) \equiv \text{Tr}\{\rho(\log(\rho) - \log(\sigma))\}. \tag{11.102}$$

Similar to the classical case, we can intuitively think of it as a distance measure between quantum states. But it is not strictly a distance measure in the mathematical sense because it is not symmetric and does not obey a triangle inequality. Nevertheless, the quantum relative entropy is always non-negative.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Theorem 11.8.1 (Positivity of Quantum Relative Entropy)**.** *The relative entropy $D(\rho \,||\, \sigma)$ is positive for any two density operators ρ and σ:*

$$D(\rho \,||\, \sigma) \geq 0. \tag{11.103}$$

Proof. Consider a spectral decomposition for ρ and σ:

$$\rho = \sum_x p(x)|\phi_x\rangle\langle\phi_x|, \qquad \sigma = \sum_y q(y)|\psi_y\rangle\langle\psi_y|, \tag{11.104}$$

where $\{|\phi_x\rangle\}$ and $\{|\psi_y\rangle\}$ are generally different orthonormal bases. Then we explicitly evaluate the formula in (11.102) for the quantum relative entropy:

$$\begin{aligned}
&D(\rho \,||\, \sigma) \\
&= \operatorname{Tr}\left\{ \sum_x p(x)|\phi_x\rangle\langle\phi_x| \log\left(\sum_{x'} p(x')|\phi_{x'}\rangle\langle\phi_{x'}| \right) \right\} \\
&\quad - \operatorname{Tr}\left\{ \sum_x p(x)|\phi_x\rangle\langle\phi_x| \log\left(\sum_y q(y)|\psi_y\rangle\langle\psi_y| \right) \right\} && (11.105) \\
&= \operatorname{Tr}\left\{ \sum_x p(x)|\phi_x\rangle\langle\phi_x| \left(\sum_{x'} \log(p(x'))|\phi_{x'}\rangle\langle\phi_{x'}| \right) \right\} \\
&\quad - \operatorname{Tr}\left\{ \sum_x p(x)|\phi_x\rangle\langle\phi_x| \sum_y \log(q(y))|\psi_y\rangle\langle\psi_y| \right\} && (11.106) \\
&= \sum_x p(x)\log(p(x)) - \sum_x p(x) \sum_y |\langle\phi_x|\psi_y\rangle|^2 \log(q(y)) && (11.107) \\
&\geq \sum_x p(x)\log(p(x)) - \sum_x p(x)\log(r(x)) && (11.108) \\
&= \sum_x p(x)\log\left(\frac{p(x)}{r(x)} \right) && (11.109) \\
&\geq 0. && (11.110)
\end{aligned}$$

The first equality follows by a direct substitution. The second equality follows because $f(A) = \sum_i f(a_i)|i\rangle\langle i|$ for any Hermitian operator A with spectral decomposition $\sum_i a_i|i\rangle\langle i|$. The third equality follows by evaluating the trace. Note that the quantity $|\langle\phi_x|\psi_y\rangle|^2$ sums to one if we sum either over x or over y—thus, we can think of it either as a conditional distribution $p(x|y)$ or $p(y|x)$, respectively. The first inequality follows by viewing the probabilities $|\langle\phi_x|\psi_y\rangle|^2$ as conditional probabilities $p(y|x)$, by noting that $-\log(z)$ is a convex function of z, and by defining $r(x) \equiv \sum_y |\langle\phi_x|\psi_y\rangle|^2 q(y)$. Note that $r(x)$ is a probability distribution because we can think of $|\langle\phi_x|\psi_y\rangle|^2$ as conditional probabilities $p(x|y)$. The fourth equality follows by collecting the logarithms, and the last inequality follows because the classical relative entropy is positive (Theorem 10.7.1). □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Corollary 11.8.1 (Subadditivity of Quantum Entropy)**.** *The von Neumann entropy is subadditive for a bipartite state ρ^{AB}:*

$$H(A)_\rho + H(B)_\rho \geq H(AB)_\rho. \tag{11.111}$$

Proof. Subadditivity of entropy is equivalent to positivity of quantum mutual information. We can prove positivity by exploiting the result of Exercise 11.8.2 and positivity of quantum relative entropy. □

The quantum relative entropy can sometimes be infinite. We consider a simple qubit example to illustrate this property and then generalize it from there. Suppose we would like to determine the quantum relative entropy between a pure state $|\psi\rangle$ and a state σ that is the following mixture:

$$\sigma \equiv \epsilon|\psi\rangle\langle\psi| + (1-\epsilon)\big|\psi^\perp\big\rangle\big\langle\psi^\perp\big|. \tag{11.112}$$

The states $|\psi\rangle$ and σ are ϵ-away from being orthogonal to each other, in the sense that:

$$\langle\psi|\sigma|\psi\rangle = \epsilon. \tag{11.113}$$

Then they are approximately distinguishable by a measurement and we would expect the relative entropy between them to be quite high. We calculate $D(|\psi\rangle \,||\, \sigma)$:

$$\begin{aligned} D(|\psi\rangle \,||\, \sigma) &= -H(|\psi\rangle) - \operatorname{Tr}\{|\psi\rangle\langle\psi| \log \sigma\} && (11.114)\\ &= -\operatorname{Tr}\big\{|\psi\rangle\langle\psi|\big(\log\epsilon|\psi\rangle\langle\psi| + \log(1-\epsilon)\big|\psi^\perp\big\rangle\big\langle\psi^\perp\big|\big)\big\} && (11.115)\\ &= -\log\epsilon. && (11.116)\end{aligned}$$

Then, the quantum relative entropy can become infinite in the limit as $\epsilon \to 0$ because

$$\lim_{\epsilon\to 0} -\log\epsilon = +\infty. \tag{11.117}$$

We generalize the above example with the following property of quantum relative entropy:

Property 11.8.1 (Infinite Quantum Relative Entropy) Suppose the support of ρ and the orthogonal support of σ have non-trivial intersection:

$$\operatorname{supp}(\rho) \cap \operatorname{supp}(\sigma)^\perp \neq \emptyset. \tag{11.118}$$

Then the quantum relative entropy is infinite:

$$D(\rho \,||\, \sigma) = +\infty. \tag{11.119}$$

We can prove this property by a simple generalization of the above "qubit" argument where we consider the following mixture and take the limit of the quantum relative entropy as $\epsilon \to 0$:

$$\epsilon\sigma + (1-\epsilon)\sigma^\perp, \tag{11.120}$$

where $\sigma^\perp$ is a strictly positive density operator that lives on the orthogonal support of σ.

We give some intuition for the latter condition in the above property. It occurs in the case where two states have orthogonal support, and there is always a measurement that can perfectly distinguish the two states in this case.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 11.8.1 Show that the following identity holds:

$$\log\bigl(\rho^A \otimes \rho^B\bigr) = \log\bigl(\rho^A\bigr) \otimes I^B + I^A \otimes \log\bigl(\rho^B\bigr). \tag{11.121}$$

Exercise 11.8.2 Show that the following identity holds:

$$D\bigl(\rho^{AB} \,||\, \rho^A \otimes \rho^B\bigr) = I(A;B)_{\rho^{AB}}. \tag{11.122}$$

Exercise 11.8.3 Show that the following identity holds:

$$D\bigl(\rho^{AB} \,||\, I^A \otimes \rho^B\bigr) = -H(A|B).$$

Exercise 11.8.4 Show that the relative entropy is invariant under unitary operations:

$$D(\rho \,||\, \sigma) = D\bigl(U\rho U^\dagger \,||\, U\sigma U^\dagger\bigr). \tag{11.123}$$

Exercise 11.8.5 (Additivity of Quantum Relative Entropy) Show that the quantum relative entropy is additive for tensor product states:

$$D(\rho_1 \otimes \rho_2 \,||\, \sigma_1 \otimes \sigma_2) = D(\rho_1 \,||\, \sigma_1) + D(\rho_2 \,||\, \sigma_2). \tag{11.124}$$

Apply the above additivity relation inductively to conclude that

$$D\bigl(\rho^{\otimes n} \,||\, \sigma^{\otimes n}\bigr) = nD(\rho \,||\, \sigma). \tag{11.125}$$

Exercise 11.8.6 (Quantum Relative Entropy of Classical-Quantum States) Show that the quantum relative entropy between classical-quantum states ρ^{XB} and σ^{XB} is as follows:

$$D\bigl(\rho^{XB} \,||\, \sigma^{XB}\bigr) = \sum_x p_X(x) D(\rho_x \,||\, \sigma_x), \tag{11.126}$$

where

$$\rho^{XB} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes \rho_x^B, \qquad \sigma^{XB} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes \sigma_x^B. \tag{11.127}$$

11.9 Quantum Information Inequalities

11.9.1 The Fundamental Quantum Information Inequality

The most fundamental information inequality in quantum information theory is the monotonicity of quantum relative entropy. The physical interpretation of this inequality is that states become less distinguishable when noise acts on them.

Theorem 11.9.1 (Monotonicity of Quantum Relative Entropy)**.** *The quantum relative entropy between two states ρ and σ can only decrease if we apply the same noisy map $\mathcal{N}$ to each state:*

$$D(\rho \,||\, \sigma) \geq D(\mathcal{N}(\rho) \,||\, \mathcal{N}(\sigma)). \tag{11.128}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Proof. We can realize any noisy map $\mathcal{N}$ by appending a state $|0\rangle^E$ to the system, applying some unitary U on the larger Hilbert space, and tracing out the environment system. With this in mind, consider the following chain of inequalities:

$$D(\rho \,||\, \sigma) = D(\rho \,||\, \sigma) + D\Big(|0\rangle\langle 0|^E \,||\, |0\rangle\langle 0|^E\Big) \tag{11.129}$$

$$= D\Big(\rho \otimes |0\rangle\langle 0|^E \,||\, \sigma \otimes |0\rangle\langle 0|^E\Big) \tag{11.130}$$

$$= D\Big(U\Big(\rho \otimes |0\rangle\langle 0|^E\Big)U^\dagger \,||\, U\Big(\sigma \otimes |0\rangle\langle 0|^E\Big)U^\dagger\Big) \tag{11.131}$$

$$\geq D(\mathcal{N}(\rho) \,||\, \mathcal{N}(\sigma)) \tag{11.132}$$

The first equality follows because the quantum relative entropy

$$D\Big(|0\rangle\langle 0|^E \,||\, |0\rangle\langle 0|^E\Big) \tag{11.133}$$

vanishes. The second equality follows from additivity of quantum relative entropy over tensor product states (see Exercise 11.8.5). The third equality follows because the quantum relative entropy is invariant under unitaries (see Exercise 11.8.4). The last inequality follows from the following simpler form of monotonicity:

$$D\big(\rho^{AB} \,||\, \sigma^{AB}\big) \geq D\big(\rho^A \,||\, \sigma^A\big). \tag{11.134}$$

The proof of (11.134) is rather involved, exploiting ideas from operator convex functions, and we prove it in full in Appendix B. □

11.9.2 Corollaries of the Fundamental Quantum Information Inequality

Monotonicity of quantum relative entropy has as its corollaries many of the important information inequalities in quantum information theory.

Corollary 11.9.1 (Strong Subadditivity)**.** *The von Neumann entropy for any tripartite state ρ^{ABC} is strongly subadditive:*

$$H(AB)_\rho + H(BC)_\rho \geq H(ABC)_\rho + H(B)_\rho. \tag{11.135}$$

Proof. Consider that

$$D\big(\rho^{ABC} \,||\, \rho^A \otimes \rho^{BC}\big) = I(A;BC)_\rho$$

The first equality follows from the result of Exercise 11.8.2. A similar relation applies for the state ρ^{AB}:

$$D\big(\rho^{AB} \,||\, \rho^A \otimes \rho^B\big) = I(A;B)_\rho. \tag{11.136}$$

Then

$$D\big(\rho^{ABC} \,||\, \rho^A \otimes \rho^{BC}\big) \geq D\big(\rho^{AB} \,||\, \rho^A \otimes \rho^B\big) \tag{11.137}$$

$$\therefore \quad I(A;BC)_\rho \geq I(A;B)_\rho \tag{11.138}$$

$$\therefore \quad I(A;B|C)_\rho \geq 0 \tag{11.139}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first line follows from monotonicity of quantum relative entropy (tracing out the C system). The second line follows by using the above results. The final line follows from the chain rule for quantum mutual information. The last line is equivalent to the statement of strong subadditivity by the result of Exercise 11.7.6. Thus, strong subadditivity follows from monotonicity of quantum relative entropy. □

Corollary 11.9.2 (Joint Convexity of Quantum Relative Entropy)**.** *The quantum relative entropy is jointly convex in its arguments:*

$$D(\rho \,||\, \sigma) \leq \sum_x p_X(x) D(\rho_x \,||\, \sigma_x), \tag{11.140}$$

where $\rho \equiv \sum_x p_X(x)\rho_x$ *and* $\sigma \equiv \sum_x p_X(x)\sigma_x$.

Proof. Consider classical-quantum states of the following form:

$$\rho^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^B, \tag{11.141}$$

$$\sigma^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \sigma_x^B. \tag{11.142}$$

Then the following chain of inequalities holds

$$\sum_x p_X(x) D(\rho_x \,||\, \sigma_x) = D\big(\rho^{XB} \,||\, \sigma^{XB}\big) \tag{11.143}$$

$$\geq D\big(\rho^B \,||\, \sigma^B\big). \tag{11.144}$$

The first equality follows from the result of Exercise 11.8.6, and the inequality follows from monotonicity of quantum relative entropy. □

Corollary 11.9.3 (Complete dephasing increases entropy)**.** *Suppose that we completely dephase a density operator* ρ *with respect to some dephasing basis* $\{|y\rangle\}$*. Let* σ *denote the dephased version of* ρ*:*

$$\sigma \equiv \Delta_Y(\rho) = \sum_y |y\rangle\langle y|\rho|y\rangle\langle y|. \tag{11.145}$$

Then the entropy $H(\sigma)$ *of the completely dephased state is greater than the entropy* $H(\rho)$ *of the original state:*

$$H(\sigma) \geq H(\rho). \tag{11.146}$$

Proof. Suppose that ρ has the following spectral decomposition:

$$\rho = \sum_x p_X(x)|x\rangle\langle x|^X. \tag{11.147}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Then the completely dephased state σ admits the following representation:

$$\sigma = \sum_y |y\rangle\langle y|\rho|y\rangle\langle y| \tag{11.148}$$

$$= \sum_y |y\rangle\langle y| \left(\sum_x p_X(x) |x\rangle\langle x|^X \right) |y\rangle\langle y| \tag{11.149}$$

$$= \sum_{x,y} p_X(x) |\langle y|x\rangle|^2 |y\rangle\langle y| \tag{11.150}$$

$$= \sum_{x,y} p_X(x) p_{Y|X}(y|x) |y\rangle\langle y|, \tag{11.151}$$

where we define $p_{Y|X}(y|x) \equiv |\langle y|x\rangle|^2$. In particular, the eigenvalues of σ are $p_Y(y) \equiv \sum_x p_X(x) p_{Y|X}(y|x)$ because σ is diagonal in the dephasing basis $\{|y\rangle\}$. We can then exploit positivity of relative entropy to obtain the following chain of inequalities:

$$0 \leq D(\rho||\sigma) \tag{11.152}$$

$$= -H(\rho) - \text{Tr}\{\rho \log \sigma\} \tag{11.153}$$

$$= -H(\rho) - \text{Tr}\left\{ \sum_x p_X(x) |x\rangle\langle x|^X \log \left(\sum_y \left(\sum_{x'} p_X(x') p_{Y|X}(y|x') \right) |y\rangle\langle y| \right) \right\} \tag{11.154}$$

$$= -H(\rho) - \text{Tr}\left\{ \sum_{x,y} p_X(x) |x\rangle\langle x|^X \log(p_Y(y)) |y\rangle\langle y| \right\} \tag{11.155}$$

The first inequality follows from positivity of quantum relative entropy. The first equality follows from the definition of quantum relative entropy. The second equality follows by expanding the term $-\text{Tr}\{\rho \log \sigma\}$. The third equality follows by evaluating the logarithm on the eigenvalues in a spectral decomposition. Continuing,

$$= -H(\rho) - \sum_{x,y} p_X(x) |\langle y|x\rangle|^2 \log(p_Y(y)) \tag{11.156}$$

$$= -H(\rho) - \sum_y p_Y(y) \log(p_Y(y)) \tag{11.157}$$

$$= -H(\rho) + H(\sigma). \tag{11.158}$$

The first equality follows from linearity of trace. The second equality follows from the definition of $p_Y(y)$. The final equality follows because the eigenvalues of σ are $p_Y(y)$. □

The quantum relative entropy itself is not equivalent to a distance measure, but it actually gives a useful upper bound on the trace distance between two quantum states. Thus, in this sense, we can think of it as being nearly equivalent to a distance measure—if the quantum relative entropy between two quantum states is small, then their trace distance is small as well.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Theorem 11.9.2 (Quantum Pinsker Inequality)**.** *The quantum relative entropy $D(\rho||\sigma)$ is an upper bound on the trace distance $\|\rho-\sigma\|_1$:*

$$\frac{1}{2\ln 2}(\|\rho-\sigma\|_1)^2 \leq D(\rho||\sigma). \tag{11.159}$$

Proof. We first prove the inequality for qubit states ρ and σ that are diagonal in the same basis:

$$\rho \equiv p|0\rangle\langle 0| + (1-p)|1\rangle\langle 1|, \tag{11.160}$$
$$\sigma \equiv q|0\rangle\langle 0| + (1-q)|1\rangle\langle 1|, \tag{11.161}$$

where $p \geq q$. This corresponds to demonstrating the following inequality:

$$p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \geq \frac{4}{2\ln 2}(p-q)^2. \tag{11.162}$$

Consider the function $g(p,q)$ where

$$g(p,q) \equiv p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) - \frac{4}{2\ln 2}(p-q)^2, \tag{11.163}$$

so that $g(p,q)$ corresponds to the difference of the LHS and the RHS in (11.162). Then

$$\frac{\partial g(p,q)}{\partial q} = -\frac{p}{q\ln 2} + \frac{1-p}{(1-q)\ln 2} - \frac{4}{\ln 2}(q-p) \tag{11.164}$$
$$= -\frac{p(1-q)}{q(1-q)\ln 2} + \frac{q(1-p)}{q(1-q)\ln 2} - \frac{4}{\ln 2}(q-p) \tag{11.165}$$
$$= \frac{q-p}{q(1-q)\ln 2} - \frac{4}{\ln 2}(q-p) \tag{11.166}$$
$$= \frac{(q-p)(4q^2-4q+1)}{q(1-q)\ln 2} \tag{11.167}$$
$$= \frac{(q-p)(2q-1)^2}{q(1-q)\ln 2} \tag{11.168}$$
$$\leq 0, \tag{11.169}$$

with the last step holding from the assumption that $p \geq q$ and $1 \geq q \geq 0$. Also, observe that both $\partial g(p,q)/\partial q = 0$ and $g(p,q) = 0$ when $p = q$. Thus, the function $g(p,q)$ is decreasing in q for every p whenever $q \leq p$ and reaches a minimum when $q = p$. So $g(p,q) \geq 0$ whenever $p \geq q$. The theorem also holds in general by applying it to $p' \equiv 1-p$ and $q' \equiv 1-q$ so that $q' \geq p'$. Now we prove the "fully quantum" version of the theorem for arbitrary states ρ and σ by exploiting the above result. Consider the projector Π onto the positive eigenspace of $\rho-\sigma$ and $I-\Pi$ is the projector onto the negative eigenspace (recall from Lemma 9.1.1 that

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$2\text{Tr}\{\Pi(\rho-\sigma)\} = \|\rho-\sigma\|_1$). Let $\mathcal{M}$ be a quantum operation that performs this projective measurement, so that

$$\mathcal{M}(\rho) = \text{Tr}\{\Pi\rho\}|0\rangle\langle 0| + \text{Tr}\{(I-\Pi)\rho\}|1\rangle\langle 1|, \tag{11.170}$$
$$\mathcal{M}(\sigma) = \text{Tr}\{\Pi\sigma\}|0\rangle\langle 0| + \text{Tr}\{(I-\Pi)\sigma\}|1\rangle\langle 1|. \tag{11.171}$$

Let $p = \text{Tr}\{\Pi\rho\}$ and $q = \text{Tr}\{\Pi\sigma\}$. Applying monotonicity of quantum relative entropy (Theorem 11.9.1) and the proof for binary variables above gives that

$$\begin{aligned} D(\rho \,||\, \sigma) &\geq D(\mathcal{M}(\rho) \,||\, \mathcal{M}(\sigma)) && (11.172)\\ &\geq \frac{4}{2\ln 2}(p-q)^2 && (11.173)\\ &= \frac{4}{2\ln 2}(\text{Tr}\{\Pi\rho\} - \text{Tr}\{\Pi\sigma\})^2 && (11.174)\\ &= \frac{1}{2\ln 2}(2\text{Tr}\{\Pi(\rho-\sigma)\})^2 && (11.175)\\ &= \frac{1}{2\ln 2}(\|\rho-\sigma\|_1)^2. && (11.176) \end{aligned}$$

□

11.9.3 The Quantum Data Processing Inequality

The quantum data processing inequality is similar in spirit to the classical data processing inequality. Recall that the classical data processing inequality states that processing classical data reduces classical correlations. The quantum data processing inequality states that processing *quantum* data reduces *quantum* correlations.

It applies to the following scenario. Suppose that Alice and Bob share some pure bipartite state $|\phi\rangle^{AB}$. The coherent information $I(A\rangle B)_\phi$ quantifies the quantum correlations present in this state. Bob then processes his system B according to some CPTP map $\mathcal{N}_1^{B\to B_1}$ to produce some quantum system B_1 and let ρ^{AB_1} denote the resulting state (in general, it could be a mixed state). He further processes his system B_1 according to some CPTP map $\mathcal{N}_2^{B_1\to B_2}$ to produce some quantum system B_2 and let σ^{AB_2} denote the state resulting from the second map. The quantum data processing inequality states that each step of quantum data processing reduces quantum correlations, in the sense that

$$I(A\rangle B)_\phi \geq I(A\rangle B_1)_\rho \geq I(A\rangle B_2)_\sigma. \tag{11.177}$$

Figure 11.1(a) depicts the scenario described above corresponding to the quantum data processing inequality. Figure 11.1(b) depicts this same scenario with the isometric extensions of the respective maps $\mathcal{N}_1^{B\to B_1}$ and $\mathcal{N}_2^{B_1\to B_2}$—this latter depiction is useful in the proof of the quantum data processing inequality.

A condition similar to the Markov condition holds for the quantum case. Each of the maps $\mathcal{N}_1^{B\to B_1}$ and $\mathcal{N}_2^{B_1\to B_2}$ acts only on one of Bob's systems—it does not act in any way on Alice's system. This behavior is what allows us to prove the quantum data processing inequality.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

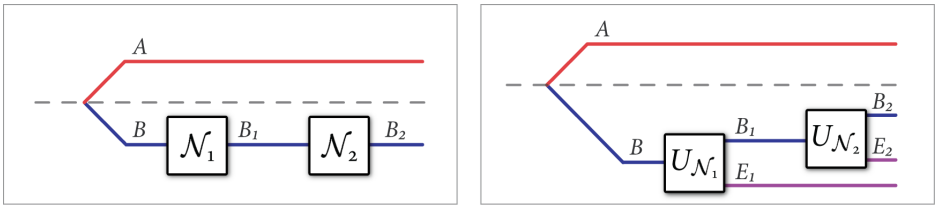

Figure 11.1: Two slightly different depictions of the quantum data processing inequality. (a) Alice and Bob begin by sharing some pure state $|\phi\rangle^{AB}$. Bob processes his system B with the CPTP map $\mathcal{N}_1$ and further processes B_1 with the CPTP map $\mathcal{N}_2$. Quantum correlations can only decrease after quantum data processing, in the sense that $I(A\rangle B) \geq I(A\rangle B_1) \geq I(A\rangle B_2)$. It also follows that $I(A;B) \geq I(A;B_1) \geq I(A;B_2)$. (b) The depiction on the right is the same as that on the left except we consider the respective isometric extensions $U_{\mathcal{N}_1}$ and $U_{\mathcal{N}_2}$ of the channels $\mathcal{N}_1$ and $\mathcal{N}_2$. The quantum state after $U_{\mathcal{N}_1}$ is a pure state shared among the systems A, B_1, and E_1, and the state after $U_{\mathcal{N}_2}$ is a pure state shared among the systems A, B_2, E_2, and E_1.

Theorem 11.9.3 (Quantum Data Processing Inequality for Coherent Information)**.** *Suppose that* $\rho^{AB_1} \equiv \mathcal{N}_1^{B\to B_1}(\phi^{AB})$ *and* $\sigma^{AB_2} \equiv \mathcal{N}_2^{B_1\to B_2}(\rho^{AB_1})$. *Then the following quantum data processing inequality applies for coherent information:*

$$I(A\rangle B)_\phi \geq I(A\rangle B_1)_\rho \geq I(A\rangle B_2)_\sigma. \tag{11.178}$$

Proof. The proof exploits the depiction of quantum data processing in Figure 11.1(b), the equivalence of the marginal entropies of a pure bipartite state (Theorem 11.2.1), and strong subadditivity (Theorem 11.7.1). First consider that

$$\begin{aligned} I(A\rangle B)_\phi &= H(B)_\phi - H(AB)_\phi && (11.179)\\ &= H(B)_\phi && (11.180)\\ &= H(A)_\phi. && (11.181) \end{aligned}$$

The first equality follows by definition, the second equality follows because the entropy of a pure state vanishes, and the third equality follows from Theorem 11.2.1. Let $|\psi\rangle^{AB_1E_1}$ be the output of the isometry $U_{\mathcal{N}_1}^{B\to B_1E_1}$:

$$|\psi\rangle^{AB_1E_1} \equiv U_{\mathcal{N}_1}^{B\to B_1E_1}|\phi\rangle^{AB}. \tag{11.182}$$

(It is also a purification of the state ρ^{AB_1}). Consider that

$$\begin{aligned} I(A\rangle B_1)_\rho &= I(A\rangle B_1)_\psi && (11.183)\\ &= H(B_1)_\psi - H(AB_1)_\psi && (11.184)\\ &= H(AE_1)_\psi - H(E_1)_\psi && (11.185) \end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Note that $H(A)_\phi = H(A)_\psi$ because no processing occurs on system A. Recall from Theorem 11.6.1 that quantum mutual information is always positive. Then the following chain of inequalities proves the first inequality in Theorem 11.9.3:

$$I(A;E_1)_\psi \geq 0 \tag{11.186}$$

$$\therefore \quad H(A)_\psi \geq H(AE_1)_\psi - H(E_1)_\psi \tag{11.187}$$

$$\therefore \quad I(A\rangle B)_\phi \geq I(A\rangle B_1)_\rho. \tag{11.188}$$

The first line applies positivity of quantum mutual information. The second line applies the definition of quantum mutual information, and the third line applies the results in (11.181) and (11.185). We prove the other inequality by a similar line of reasoning (though this time we resort to the positivity of conditional mutual information in Theorem 11.7.1). Let $|\varphi\rangle^{AB_2E_2E_1}$ be the output of the isometry $U_{\mathcal{N}_2}^{B_1\to B_2E_2}$:

$$|\varphi\rangle^{AB_2E_2E_1} \equiv U_{\mathcal{N}_2}^{B_1\to B_2E_2}|\psi\rangle^{AB_1E_1}. \tag{11.189}$$

(It is also a purification of the state σ^{AB_2}). Then

$$I(A\rangle B_2)_\sigma = I(A\rangle B_2)_\varphi \tag{11.190}$$

$$= H(B_2)_\varphi - H(AB_2)_\varphi \tag{11.191}$$

$$= H(AE_1E_2)_\varphi - H(E_1E_2)_\varphi \tag{11.192}$$

$$= H(AE_2|E_1)_\varphi - H(E_2|E_1)_\varphi \tag{11.193}$$

The third equality follows from Theorem 11.2.1, and the last follows by adding and subtracting the marginal entropy $H(E_1)_\varphi$ and recalling the definition of conditional quantum entropy. Also,

$$I(A\rangle B_1)_\rho = H(AE_1)_\psi - H(E_1)_\psi \tag{11.194}$$

$$= H(AE_1)_\varphi - H(E_1)_\varphi \tag{11.195}$$

$$= H(A|E_1)_\varphi \tag{11.196}$$

The first equality follows from (11.185). The second equality follows because there is no quantum processing on systems A and E_1 to produce state φ from state ψ. The third equality follows from the definition of conditional quantum entropy. Recall from Theorem 11.7.1 that conditional quantum mutual information is always positive. Then the following chain of inequalities proves the second inequality in Theorem 11.9.3:

$$I(A;E_2|E_1)_\varphi \geq 0 \tag{11.197}$$

$$\therefore \quad H(A|E_1)_\varphi \geq H(AE_2|E_1)_\varphi - H(E_2|E_1)_\varphi \tag{11.198}$$

$$\therefore \quad I(A\rangle B_1)_\rho \geq I(A\rangle B_2)_\sigma. \tag{11.199}$$

The first line applies positivity of conditional quantum mutual information. The second line applies the definition of conditional quantum mutual information, and the third line applies the result in (11.193) and (11.196). □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Corollary 11.9.4 (Quantum Data Processing Inequality for Quantum Mutual Information)**.** *Suppose that* $\rho^{AB_1} \equiv \mathcal{N}_1^{B\to B_1}(\phi^{AB})$ *and* $\sigma^{AB_2} \equiv \mathcal{N}_2^{B_1\to B_2}(\rho^{AB_1})$*. Then the following quantum data processing inequality applies to the quantum mutual information:*

$$I(A;B)_\phi \geq I(A;B_1)_\rho \geq I(A;B_2)_\sigma. \tag{11.200}$$

That is, we have

$$I(A;B) \geq I(A_1;B) \geq I(A_2;B), \tag{11.201}$$

for some maps applied to the A *system in the order* $A \to A_1 \to A_2$*.*

Proof. The proof follows because the quantum mutual information $I(A;B) = H(A)+I(A\rangle B)$ and we can apply the quantum data processing inequality for coherent information. Though, the quantum data processing inequality is symmetric for the case of quantum mutual information (QMI) because the QMI itself is symmetric. □

Exercise 11.9.1 (Holevo Bound) Use the quantum data processing inequality to show that the Holevo information $\chi(\mathcal{E})$ is an upper bound on the accessible information $I_{\text{acc}}(\mathcal{E})$:

$$I_{\text{acc}}(\mathcal{E}) \leq \chi(\mathcal{E}). \tag{11.202}$$

Exercise 11.9.2 (Shannon Entropy versus von Neumann Entropy of an Ensemble) Consider an ensemble $\{p_X(x), |\psi_x\rangle\}$. The expected density operator of the ensemble is

$$\rho \equiv \sum_x p_X(x)|\psi_x\rangle\langle\psi_x|. \tag{11.203}$$

Use the quantum data processing inequality to show that the Shannon entropy $H(X)$ is never less than the von Neumann entropy of the expected density operator ρ:

$$H(X) \geq H(\rho). \tag{11.204}$$

(Hint: Begin with a classical common randomness state $\sum_x p_X(x)|x\rangle\langle x|^X \otimes |x\rangle\langle x|^{X'}$ and apply a preparation map to system X'). Conclude that the Shannon entropy of the ensemble is strictly greater than the von Neumann entropy whenever the states in the ensemble are non-orthogonal.

Exercise 11.9.3 Use the idea in the above exercise to show that the conditional entropy $H(X|B)_\rho$ is always non-negative whenever the state ρ^{XB} is a classical-quantum state:

$$\rho^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^B. \tag{11.205}$$

Exercise 11.9.4 (Separability and Negativity of Coherent Information) Show that the following inequality holds for any separable state ρ^{AB}:

$$\max\Big\{I(A\rangle B)_{\rho^{AB}}, I(B\rangle A)_{\rho^{AB}}\Big\} \leq 0. \tag{11.206}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

11.9.4 Continuity of Quantum Entropy

Suppose that two density operators ρ and σ are close in trace distance. We might then expect several properties to hold: the fidelity between them should be close to one and their entropies should be close. Theorem 9.3.1 states that the fidelity is close to one if the trace distance is small.

An important theorem below, the Alicki-Fannes' inequality, states that conditional quantum entropies are close as well. This theorem usually finds application in a proof of a converse theorem in quantum Shannon theory. Usually, the specification of any good protocol (in the sense of asymptotically vanishing error) involves placing a bound on the trace distance between the actual state resulting from a protocol and the ideal state that it should produce. The Alicki-Fannes' inequality then allows us to translate these statements of error into informational statements that bound the asymptotic rates of communication in any good protocol. We give the full proof of this theorem below, and ask you to prove variations of it in the exercises below.

Theorem 11.9.4 (Alicki-Fannes Inequality)**.** *For any states ρ^{AB} and σ^{AB} with $\left\|\rho^{AB} - \sigma^{AB}\right\|_1 \leq \epsilon$,*

$$|H(A|B)_\rho - H(A|B)_\sigma| \leq 4\epsilon \log d_A + 2H_2(\epsilon), \tag{11.207}$$

where $H_2(\epsilon)$ is the binary entropy function.

Proof. Suppose that $\left\|\rho^{AB} - \sigma^{AB}\right\|_1 = \epsilon$ and $\epsilon < 1$ (we are really only concerned with small ϵ). Let $\tilde{\rho}^{AB}$ and $\tilde{\sigma}^{AB}$ denote the following density operators:

$$\tilde{\rho}^{AB} \equiv \frac{1}{\epsilon}\left|\rho^{AB} - \sigma^{AB}\right|, \tag{11.208}$$

$$\tilde{\sigma}^{AB} \equiv \frac{1-\epsilon}{\epsilon}\left(\rho^{AB} - \sigma^{AB}\right) + \tilde{\rho}^{AB}. \tag{11.209}$$

(You should verify that both are indeed valid density operators!). We introduce a classical-quantum state γ^{XAB}:

$$\gamma^{XAB} \equiv (1-\epsilon)|0\rangle\langle 0|^X \otimes \rho^{AB} + \epsilon|1\rangle\langle 1|^X \otimes \tilde{\rho}^{AB}, \tag{11.210}$$

Consider that

$$\gamma^{AB} = \text{Tr}_X\left\{\gamma^{XAB}\right\} = (1-\epsilon)\rho^{AB} + \epsilon\tilde{\rho}^{AB}. \tag{11.211}$$

The following crucial equivalence holds as well:

$$\gamma^{AB} = (1-\epsilon)\sigma^{AB} + \epsilon\tilde{\sigma}^{AB}, \tag{11.212}$$

by examining the definition of $\tilde{\sigma}^{AB}$ in (11.209). Thus, we can mix the states ρ^{AB} and $\tilde{\rho}^{AB}$ and the states σ^{AB} and $\tilde{\sigma}^{AB}$ in the same proportions to get the state γ^{AB}. We now prove the following inequality:

$$\left|H(A|B)_\rho - H(A|B)_\gamma\right| \leq 2\epsilon \log d_A + H_2(\epsilon). \tag{11.213}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We first prove that $H(A|B)_\rho - H(A|B)_\gamma \leq 2\epsilon \log d_A + H_2(\epsilon)$. The following inequality holds because the conditional entropy is concave (Exercise 11.7.4):

$$H(A|B)_\gamma \geq (1-\epsilon)H(A|B)_\rho + \epsilon H(A|B)_{\tilde{\rho}}. \tag{11.214}$$

The above inequality implies the following one:

$$\begin{aligned} H(A|B)_\rho - H(A|B)_\gamma &\leq \epsilon\Big(H(A|B)_\rho - H(A|B)_{\tilde{\rho}}\Big) && (11.215)\\ &\leq 2\epsilon \log d_A && (11.216)\\ &\leq 2\epsilon \log d_A + H(\epsilon). && (11.217)\end{aligned}$$

The second inequality holds because $\log d_A$ is the largest that each of $H(A|B)_\rho$ and $H(A|B)_{\tilde{\rho}}$ can be (Theorem 11.5.1). We now prove the other bound in the absolute value in (11.213): $H(A|B)_\gamma - H(A|B)_\rho \leq 2\epsilon \log d_A + H(\epsilon)$. Concavity of quantum entropy (Exercise 11.6.9) implies the following inequality:

$$H(B)_\gamma \geq (1-\epsilon)H(B)_\rho + \epsilon H(B)_{\tilde{\rho}}. \tag{11.218}$$

Consider that the following inequality holds

$$H(AB)_\gamma \leq H(ABX)_\gamma, \tag{11.219}$$

because the addition of a classical system can only increase the entropy. Then

$$\begin{aligned} H(AB)_\gamma &\leq H(AB|X)_\gamma + H(X)_\gamma && (11.220)\\ &= (1-\epsilon)H(AB)_\rho + \epsilon H(AB)_{\tilde{\rho}} + H_2(\epsilon). && (11.221)\end{aligned}$$

Combining (11.218) and (11.221) gives the following inequality:

$$\begin{aligned} H(A|B)_\gamma &\leq (1-\epsilon)H(A|B)_\rho + \epsilon H(A|B)_{\tilde{\rho}} + H_2(\epsilon) && (11.222)\\ \Rightarrow H(A|B)_\gamma - H(A|B)_\rho &\leq \epsilon\Big(H(A|B)_{\tilde{\rho}} - H(A|B)_\rho\Big) + H_2(\epsilon) && (11.223)\\ &\leq 2\epsilon \log d_A + H_2(\epsilon). && (11.224)\end{aligned}$$

By the same line of reasoning, the following inequality holds

$$\Big|H(A|B)_\sigma - H(A|B)_\gamma\Big| \leq 2\epsilon \log d_A + H_2(\epsilon). \tag{11.225}$$

We can now complete the proof of Theorem 11.9.4. Consider the following chain of inequalities:

$$\begin{aligned} |H(A|B)_\rho - H(A|B)_\sigma| &\leq |H(A|B)_\rho - H(A|B)_\gamma| + |H(A|B)_\sigma - H(A|B)_\gamma| && (11.226)\\ &\leq 4\epsilon \log d_A + 2H_2(\epsilon). && (11.227)\end{aligned}$$

The first inequality follows from the triangle inequality and the second follows from (11.213) and (11.225). □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

A corollary of the above theorem is Fannes' inequality, which provides a better upper bound on the entropy difference between two states ρ and σ.

Theorem 11.9.5 (Fannes' Inequality)**.** *For any ρ and σ with $||\rho - \sigma||_1 \leq \epsilon$, the following inequality holds:*

$$|H(\rho) - H(\sigma)| \leq 2\epsilon \log d + 2H_2(\epsilon). \tag{11.228}$$

Exercise 11.9.5 Prove Fannes' inequality. (Hint: You can exploit the proof of the Alicki-Fannes' inequality.)

A slight (and optimal) improvement of the above is due to Audenaert and is known as the Fannes-Audenaert inequality:

Theorem 11.9.6 (Fannes-Audenaert Inequality)**.** *For any ρ and σ with $T \equiv \frac{1}{2}||\rho - \sigma||_1$, the following inequality holds:*

$$|H(\rho) - H(\sigma)| \leq T \log(d-1) + H_2(T). \tag{11.229}$$

Exercise 11.9.6 (Alicki-Fannes' inequality for Coherent Information) Prove that

$$|I(A\rangle B)_\rho - I(A\rangle B)_\sigma| \leq 4\epsilon \log d_A + 2H_2(\epsilon), \tag{11.230}$$

for any ρ^{AB} and σ^{AB} with $\left\|\rho^{AB} - \sigma^{AB}\right\|_1 \leq \epsilon$.

Exercise 11.9.7 (Alicki-Fannes' inequality for Quantum Mutual Information) Prove that

$$|I(A;B)_\rho - I(A;B)_\sigma| \leq 6\epsilon \log d_A + 4H_2(\epsilon), \tag{11.231}$$

for any ρ^{AB} and σ^{AB} with $\left\|\rho^{AB} - \sigma^{AB}\right\|_1 \leq \epsilon$.

11.9.5 The Uncertainty Principle in the Presence of Quantum Memory

The uncertainty principle reviewed in Section 3.4.2 aims to capture a fundamental feature of quantum mechanics, namely, that there is an unavoidable uncertainty in the measurement outcomes of incompatible (non-commuting) observables. This uncertainty principle is a radical departure from classical intuitions, where, in principle, it seems as if there should not be any obstacle to measuring incompatible observables such as position and momentum.

Though, the uncertainty principle that we reviewed before (the standard version in most textbooks) suffers from a few deficiencies. First, the measure of uncertainty used there is the standard deviation, which is not just a function of the probabilities of measurement outcomes but also of the values of the outcomes. Thus, the values of the outcomes may skew the uncertainty measure (though, one could always relabel the values in order to avoid this difficulty). More importantly though, from an information-theoretic perspective, there is not a clear operational interpretation for the standard deviation as there is for entropy. Second, the lower bound in (3.111) depends not only on the observables but also the state. In

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 3.4.5, we saw how this lower bound can vanish for a state even when the distributions corresponding to the measurement outcomes in fact do have uncertainty. So, it would be ideal to separate this lower bound into two terms: one which depends only on measurement incompatibility and another which depends only on the state.

Additionally, it might seem as if giving two parties access to a maximally entangled state allows them to defy the uncertainty principle (and this is what confounded Einstein, Podolsky, and Rosen after quantum mechanics had been established). Indeed, suppose that Alice and Bob share a Bell state $|\Phi^+\rangle = 2^{-1/2}(|00\rangle + |11\rangle) = 2^{-1/2}(|{+}{+}\rangle + |{-}{-}\rangle)$. If Alice measures the Pauli Z observable on her system, then Bob can guess the outcome of her measurement with certainty. Also, if Alice were instead to measure the Pauli X observable on her system, then Bob would also be able to guess the outcome of her measurement with certainty, in spite of the fact that Z and X are incompatible observables. So, a revision of the uncertainty principle is clearly needed to account for this possibility, in the scenario where Bob shares a *quantum memory* correlated with Alice's system.

The *uncertainty principle in the presence of quantum memory* is such a revision that meets all of the desiderata stated above. It quantifies uncertainty in terms of von Neumann entropy rather than with standard deviation, and it also accounts for the scenario in which an observer has a quantum memory correlated with the system being measured. So, suppose that Alice and Bob share systems A and B, respectively, that are in some state ρ^{AB}. If Alice performs a POVM $\{\Lambda_x\}$ on her system A, then the post-measurement state is as follows:

$$\sigma^{XB} \equiv \sum_x |x\rangle\langle x|^X \otimes \text{Tr}_A\{(\Lambda_x^A \otimes I^B)\rho^{AB}\}. \tag{11.232}$$

In the above classical-quantum state, the measurement outcomes x are encoded into orthonormal states $\{|x\rangle\}$ of the classical register X, and the probability for obtaining outcome x is $\text{Tr}\{(\Lambda_x^A \otimes I^B)\rho^{AB}\}$. We would like to quantify the uncertainty that Bob has about the outcome of the measurement, and a natural quantity for doing so is the conditional quantum entropy $H(X|B)_\sigma$. Similarly, starting from the state ρ^{AB}, Alice could choose to measure some other POVM $\{\Gamma_z\}$ on her system A. In this case, the post-measurement state is as follows:

$$\tau^{ZB} \equiv \sum_z |z\rangle\langle z|^Z \otimes \text{Tr}_A\{(\Gamma_z^A \otimes I^B)\rho^{AB}\}, \tag{11.233}$$

with a similar interpretation as before. We could also quantify Bob's uncertainty about the measurement outcome z in terms of the conditional quantum entropy $H(Z|B)_\tau$. We define Bob's total uncertainty about the measurements to be the sum of both entropies: $H(X|B)_\sigma + H(Z|B)_\tau$. We will call this the *uncertainty sum*, in analogy with the uncertainty product in (3.111).

We stated above that it would be desirable to have a lower bound on the uncertainty sum consisting of a measurement incompability term and a state-dependent term. One way to quantify the incompatibility for the POVMs $\{\Lambda_x\}$ and $\{\Gamma_z\}$ is in terms of the following quantity:

$$c \equiv \max_{x,z} \left\| \sqrt{\Lambda_x}\sqrt{\Gamma_z} \right\|_\infty^2, \tag{11.234}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $\|\cdot\|_\infty$ is the infinity norm of an operator (for the finite-dimensional case, $\|A\|_\infty$ is just the maximal eigenvalue of $|A|$). To grasp an intuition for this incompatibility measure, suppose that $\{\Lambda_x\}$ and $\{\Gamma_z\}$ are actually von Neumann measurements with one common element. In this case, it follows that $c = 1$, so that the measurements are regarded as maximally compatible. On the other hand, if the measurements are of Pauli observables X and Z, these are maximally incompatible for a two-dimensional Hilbert space and $c = 1/2$. We now state the uncertainty principle in the presence of quantum memory:

Theorem 11.9.7 (Uncertainty Principle with Quantum Memory)**.** *Suppose that Alice and Bob share a state ρ^{AB} and that Alice performs either of the POVMs $\{\Lambda_x\}$ or $\{\Gamma_z\}$ on her share of the state (with at least one of $\{\Lambda_x\}$ or $\{\Gamma_z\}$ being a rank-one POVM). Then Bob's total uncertainty about the measurement outcomes has the following lower bound:*

$$H(X|B)_\sigma + H(Z|B)_\tau \geq \log_2(1/c) + H(A|B)_\rho, \tag{11.235}$$

where the states σ^{XB} and τ^{ZB} are defined in (11.232) and (11.233), respectively, and the measurement incompatibility is defined in (11.234).

Interestingly, the lower bound given in the above theorem consists of both the measurement incompatibility and the state-dependent term $H(A|B)_\rho$. As we know from Exercise 11.9.4, when the conditional quantum entropy $H(A|B)_\rho$ becomes negative, this implies that the state ρ^{AB} is entangled (but not necessarily the converse). Thus, a negative conditional entropy implies that the lower bound on the uncertainty sum can become lower than $\log_2(1/c)$, and furthermore, that it might be possible to reduce Bob's total uncertainty about the measurement outcomes down to zero. Indeed, this is the case for the example we mentioned before with measurements of Pauli X and Z on the maximally entangled Bell state. One can verify for this case that $\log(1/c) = 1$ and $H(A|B) = -1$, so that this is consistent with the fact that $H(X|B)_\sigma + H(Z|B)_\tau = 0$ for this example. We now give a path to proving the above theorem (leaving the final steps as an exercise).

Proof. We actually prove the following uncertainty relation instead:

$$H(X|B)_\sigma + H(Z|E)_\omega \geq \log_2(1/c), \tag{11.236}$$

where ω^{ZE} is a classical-quantum state of the following form:

$$\omega^{ZE} \equiv \sum_z |z\rangle\langle z|^Z \otimes \mathrm{Tr}_{AB}\{(\Gamma_z^A \otimes I^{BE})\phi_\rho^{ABE}\}, \tag{11.237}$$

and ϕ_ρ^{ABE} is a purification of ρ^{AB}. We leave it as an exercise to demonstrate that the above uncertainty relation implies the one in the statement of the theorem whenever Γ_z^A is a rank-one POVM. Consider defining the following isometric extensions of the measurement maps for $\{\Lambda_x\}$ and $\{\Gamma_z\}$:

$$V_\Lambda^{A\to XX'A} \equiv \sum_x |x\rangle^X \otimes |x\rangle^{X'} \otimes \sqrt{\Lambda_x}, \tag{11.238}$$

$$V_\Gamma^{A\to ZZ'A} \equiv \sum_z |z\rangle^Z \otimes |z\rangle^{Z'} \otimes \sqrt{\Gamma_z}, \tag{11.239}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $\{|x\rangle\}$ and $\{|z\rangle\}$ are both orthonormal bases. Let $\omega^{ZZ'ABE}$ denote the following state:

$$|\omega\rangle^{ZZ'ABE} \equiv V_\Gamma^{A\to ZZ'A}|\phi_\rho\rangle^{ABE}, \tag{11.240}$$

so that $\omega^{ZE} = \mathrm{Tr}_{Z'AB}\{\omega^{ZZ'ABE}\}$. Now consider that

$$H(Z|E)_\omega = -H(Z|Z'AB)_\omega, \tag{11.241}$$

so that (11.236) is equivalent to

$$-H(Z|Z'AB)_\omega \geq \log_2(1/c) - H(X|B)_\sigma. \tag{11.242}$$

Recalling the result of Exercise 11.8.3, we then have that the above is equivalent to

$$D\Big(\omega^{ZZ'AB} \;||\; I^Z \otimes \omega^{Z'AB}\Big) \geq \log_2(1/c) + D\big(\sigma^{XB} \;||\; I^Z \otimes \sigma^B\big), \tag{11.243}$$

where we observe that $\sigma^B = \omega^B$. So we aim to prove the above inequality. Consider the following chain of inequalities:

$$D\Big(\omega^{ZZ'AB} \;||\; I^Z \otimes \omega^{Z'AB}\Big) \tag{11.244}$$

$$\geq D\Big(\omega^{ZZ'AB} \;||\; V_\Gamma V_\Gamma^\dagger\Big(I^Z \otimes \omega^{Z'AB}\Big)V_\Gamma V_\Gamma^\dagger\Big) \tag{11.245}$$

$$= D\Big(V_\Gamma^\dagger \omega^{ZZ'AB} V_\Gamma \;||\; V_\Gamma^\dagger\Big(I^Z \otimes \omega^{Z'AB}\Big)V_\Gamma\Big) \tag{11.246}$$

$$= D\Big(\rho^{AB} \;||\; V_\Gamma^\dagger\Big(I^Z \otimes \omega^{Z'AB}\Big)V_\Gamma\Big) \tag{11.247}$$

$$= D\Big(V_\Lambda \rho^{AB} V_\Lambda^\dagger \;||\; V_\Lambda V_\Gamma^\dagger\Big(I^Z \otimes \omega^{Z'AB}\Big)V_\Gamma V_\Lambda^\dagger\Big) \tag{11.248}$$

The first inequality follows from monotonicity of quantum relative entropy under the map $\rho \to \Pi\rho\Pi + (I-\Pi)\rho(I-\Pi)$, where the projector $\Pi \equiv V_\Gamma V_\Gamma^\dagger$. The first equality follows from invariance of quantum relative entropy under isometries. The second equality follows from the fact that $V_\Gamma^\dagger \omega^{ZZ'AB} V_\Gamma = \rho^{AB}$. The third equality again follows from invariance of quantum relative entropy under isometries. Let us define $\sigma^{XX'ABE}$ as

$$|\sigma\rangle^{XX'ABE} \equiv V_\Lambda^{A\to XX'A}|\phi_\rho\rangle^{ABE}. \tag{11.249}$$

We then have that the last line above is equal to

$$D\Big(\sigma^{XX'AB} \;||\; V_\Lambda V_\Gamma^\dagger\Big(I^Z \otimes \omega^{Z'AB}\Big)V_\Gamma V_\Lambda^\dagger\Big), \tag{11.250}$$

and explicitly evaluating $V_\Lambda V_\Gamma^\dagger\big(I^Z \otimes \omega^{Z'AB}\big)V_\Gamma V_\Lambda^\dagger$ as

$$V_\Lambda V_\Gamma^\dagger\Big(I^Z \otimes \omega^{Z'AB}\Big)V_\Gamma V_\Lambda^\dagger \tag{11.251}$$

$$= V_\Lambda \sum_{z',z} \langle z'|z\rangle^Z \Big(\langle z'|^{Z'} \otimes \sqrt{\Gamma_{z'}}^A\Big)\omega^{Z'AB}\Big(|z\rangle^{Z'} \otimes \sqrt{\Gamma_z}^A\Big)V_\Lambda^\dagger \tag{11.252}$$

$$= V_\Lambda \sum_{z} \Big(\langle z|^{Z'} \otimes \sqrt{\Gamma_z}^A\Big)\omega^{Z'AB}\Big(|z\rangle^{Z'} \otimes \sqrt{\Gamma_z}^A\Big)V_\Lambda^\dagger, \tag{11.253}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

gives that this is equal to

$$D\left(\sigma^{XX'AB} \;||\; V_\Lambda \sum_z \left(\langle z|^{Z'} \otimes \sqrt{\Gamma_z}^{A}\right)\omega^{Z'AB}\left(|z\rangle^{Z'} \otimes \sqrt{\Gamma_z}^{A}\right)V_\Lambda^\dagger\right). \tag{11.254}$$

We trace out the $X'A$ systems and exploit monotonicity of quantum relative entropy and cyclicity of trace to show that the above is not less than

$$D\left(\sigma^{XB} \;||\; \sum_{z,x} |x\rangle\langle x|^X \otimes \mathrm{Tr}_{Z'A}\left\{\left(|z\rangle\langle z|^{Z'} \otimes \sqrt{\Gamma_z}\Lambda_x\sqrt{\Gamma_z}^{A}\right)\omega^{Z'AB}\right\}\right). \tag{11.255}$$

Using the fact that $\sqrt{\Gamma_z}\Lambda_x\sqrt{\Gamma_z} = |\sqrt{\Gamma_z}\sqrt{\Lambda_x}|^2 \leq cI$ and $-\log$ is operator monotone, we have that the above is not less than

$$\begin{aligned} D\big(\sigma^{XB} \;||\; c\; I^X \otimes \omega^B\big) &= \log_2(1/c) + D\big(\sigma^{XB} \;||\; I^X \otimes \omega^B\big) && (11.256)\\ &= \log_2(1/c) + D\big(\sigma^{XB} \;||\; I^X \otimes \sigma^B\big), && (11.257)\end{aligned}$$

which finally proves the inequality in (11.236). We now leave it as an exercise to prove the statement of the theorem starting from the inequality in (11.236). □

Exercise 11.9.8 Prove that (11.236) implies Theorem 11.9.7.

Exercise 11.9.9 Prove that Theorem 11.9.7 implies the following entropic uncertainty relation for a state ρ^A on a single system:

$$H(X) + H(Z) \geq \log_2(1/c) + H(A)_\rho, \tag{11.258}$$

where $H(X)$ and $H(Z)$ are the Shannon entropies of the measurement outcomes.

11.10 History and Further Reading

The von Neumann entropy and its derivatives, such as the quantum conditional entropy and quantum mutual information, are useful information measures and suffice for our studies in this book. Though, the von Neumann entropy is certainly not the only information measure worthy of study. In recent years, entropic measures such as the min- and max-entropy have emerged (and their smoothed variants), and they are useful in developing a more general theory of quantum information that applies beyond the IID setting that we study in this book. In fact, one could view this theory as more fundamental than the theory presented in this book, since the "one-shot" results often imply the IID results studied in this book. Rather than developing this theory in full, we point to several excellent references on the subject [207, 177, 237, 60, 238, 64].

Fannes proved the inequality that bears his name [91], and Audenaert later gave a significant improvement of the inequality [12]. Alicki and Fannes proved the inequality in

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Theorem 11.9.4 in Ref. [9]—the proof given here is the same as their proof. The coherent information first appeared in Ref. [218], where Schumacher and Nielsen proved that it obeys a quantum data processing inequality (this was the first clue that the coherent information would be an important information quantity for the transmission of quantum data through a noisy quantum channel). Schumacher and Westmoreland proved the bound regarding quantum relative entropy and trace distance in Theorem 11.9.2 [221]. Lieb and Ruskai proved the strong subadditivity of quantum entropy [184].

Entropic uncertainty relations have a long and interesting history. We do not review this history here but instead point to the survey article [245]. After this survey appeared, there has been much interest in entropic uncertainty relations, with the most notable advance being the entropic uncertainty relation in the presence of quantum memory [38]. The proof that we give for Theorem 11.9.7 is the same as that in Ref. [56], which in turn exploits ideas from Ref. [239].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 12

The Information of Quantum Channels

We introduced several classical and quantum entropic quantities in Chapters 10 and 11: entropy, conditional entropy, joint entropy, mutual information, relative entropy, and conditional mutual information. Each of these entropic quantities is static, in the sense that each is with respect to random variables or quantum systems that certain parties possess.

In this chapter, we introduce several dynamic entropic quantities for channels, whether they be classical or quantum. We derive these measures by exploiting the static measures from the two previous chapters. We send part of a system through a channel, compute a static measure with respect to the input-output state, and maximize the static measure over all possible systems that we can transmit through the channel. This process then gives rise to a dynamic measure that quantifies the ability of a channel to preserve correlations. For example, we could send half of a pure entangled state $|\phi\rangle^{AA'}$ through a quantum channel $\mathcal{N}^{A'\to B}$—this transmission gives rise to some noisy state $\mathcal{N}^{A'\to B}(\phi^{AA'})$. We would then take the mutual information of the resulting state and maximize the mutual information over all such pure input states:

$$\max_{\phi^{AA'}} I(A;B)_{\mathcal{N}^{A'\to B}(\phi^{AA'})}. \tag{12.1}$$

The above quantity is a dynamic information measure of the channel's abilities to preserve correlations—Section 12.4 introduces this quantity as the mutual information of the channel $\mathcal{N}$.

For now, we simply think of the quantities in this chapter as measures of a channel's ability to preserve correlations. Later, we show that these quantities have explicit operational interpretations in terms of a channel's ability to perform a certain task, such as the transmission of classical or quantum information.[1] Such an operational interpretation gives meaning to an entropic measure—otherwise, it is difficult to understand a measure in an information-theoretic sense without having a specific operational task to which it corresponds.

[1] Giving operational interpretations to informational measures is in fact the main goal of this book!

Recall that the entropy obeys an additivity property for any two independent random variables X_1 and X_2:

$$H(X_1, X_2) = H(X_1) + H(X_2). \tag{12.2}$$

The above additivity property extends to a large sequence $X_1, \ldots, X_n$ of independent and identically distributed random variables. That is, applying (12.2) inductively shows that a simple formula $nH(X)$ is the entropy of the sequence:

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) = \sum_{i=1}^{n} H(X) = nH(X), \tag{12.3}$$

where random variable X has the same distribution as all of $X_1, \ldots, X_n$. Similarly, quantum entropy is additive for any two quantum systems in a product state $\rho \otimes \sigma$:

$$H(\rho \otimes \sigma) = H(\rho) + H(\sigma), \tag{12.4}$$

and applying (12.4) inductively to a sequence of quantum states gives the following similar simple formula: $H(\rho^{\otimes n}) = nH(\rho)$. Additivity is a desirable property and a natural expectation that we have for any measure of information on independent systems.

In analogy with the static measures, we would like additivity to hold for the dynamic information measures. Without additivity holding, we cannot really make sense of a given measure because we would have to evaluate the measure on a potentially infinite number of independent channel uses. This evaluation on so many channel uses is an impossible optimization problem. Additionally, the requirement to maximize over so many uses of the channel does not identify a given measure as a unique measure of a channel's ability to perform a certain task. As we see later, there could be other measures that are equal to the original one when we take the limit of many channel uses. Thus, a measure does not have much substantive meaning if additivity does not hold.

We devote this chapter to the discussion of several dynamic measures. Additivity holds in the general case for only two of the dynamic measures presented here: the mutual information of a classical channel and the mutual information of a quantum channel. For all other measures, there are known counterexamples of channels for which additivity does not hold. In this chapter, we do not discuss the counterexamples, but instead focus only on classes of channels for which additivity does hold, in an effort to understand it in a technical sense. The proof techniques for additivity exploit many of the ideas introduced in the two previous chapters and give us a chance to practice with what we have learned there on one of the most important problems in quantum Shannon theory.

12.1 Mutual Information of a Classical Channel

Suppose that we would like to determine how much information we can transmit through a classical channel $\mathcal{N}$. Recall our simplified model of a classical channel $\mathcal{N}$ from Chapter 2, in which some conditional probability density $p_{Y|X}(y|x)$ models the effects of noise. That is, we obtain some random variable Y if we input a random variable X to the channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

What is a good measure of the information throughput of this channel? The mutual information is perhaps the best starting point. Suppose that random variables X and Y are Bernoulli. If the classical channel is noiseless and X is completely random, the input and output random variables X and Y are perfectly correlated and the mutual information $I(X;Y)$ is equal to one bit, implying that the sender can transmit one bit per transmission as we would expect. If the classical channel is completely noisy (in the sense that it prepares an output that is constant irrespective of the input), the input and output random variables are independent and the mutual information is equal to zero bits. This result matches our intuition that the sender should not be able to transmit any information through this completely noisy channel.

In the above model for a classical channel, the conditional probability density $p_{Y|X}(y|x)$ remains fixed, but we can "play around" with the input random variable X by modifying its probability density $p_X(x)$.[2] Thus, we still "have room" for optimizing the mutual information of the channel $\mathcal{N}$ by modifying this input density. This gives us the following definition:

Definition 12.1.1 (Mutual Information of a Classical Channel)**.** *The mutual information $I(\mathcal{N})$ of the classical channel $\mathcal{N}$ is as follows:*

$$I(\mathcal{N}) \equiv \max_{p_X(x)} I(X;Y). \tag{12.5}$$

12.1.1 Regularization of the Mutual Information of a Classical Channel

We now consider whether exploiting multiple uses of a classical channel $\mathcal{N}$ and allowing for correlations between its inputs can increase its mutual information. That is, suppose that we have two independent uses of a classical channel $\mathcal{N}$ available. Let X_1 and X_2 denote the input random variables to the respective first and second copies of the channel, and let Y_1 and Y_2 denote the output random variables. Each of the two uses of the channel are equivalent to the mapping $p_{Y|X}(y|x)$ so that the channel uses are independent and identically distributed. Let $\mathcal{N} \otimes \mathcal{N}$ denote the *tandem channel* that corresponds to the mapping

$$p_{Y_1,Y_2|X_1,X_2}(y_1,y_2|x_1,x_2) = p_{Y_1|X_1}(y_1|x_1)p_{Y_2|X_2}(y_2|x_2), \tag{12.6}$$

where both $p_{Y_1|X_1}(y_1|x_1)$ and $p_{Y_2|X_2}(y_2|x_2)$ are equivalent to the mapping $p_{Y|X}(y|x)$. The mutual information of a classical tandem channel is as follows:

$$I(\mathcal{N} \otimes \mathcal{N}) \equiv \max_{p_{X_1,X_2}(x_1,x_2)} I(X_1,X_2;Y_1,Y_2). \tag{12.7}$$

We might think that we could increase the mutual information of this classical channel by allowing for correlations between the inputs to the channels through a correlated distribution $p_{X_1,X_2}(x_1,x_2)$. That is, there could be some superadditive effect if the mutual

[2] Recall the idea from Section 2.2.4 where Alice and Bob actually choose a code for the channel randomly according to the density $p_X(x)$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

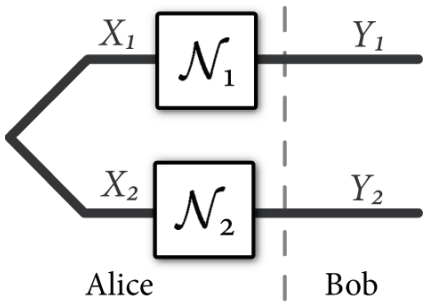

Figure 12.1: The above figure displays the scenario for determining whether the mutual information of two classical channels $\mathcal{N}_1$ and $\mathcal{N}_2$ is additive. The question of additivity is equivalent to the possibility of classical correlations being able to enhance the mutual information of two classical channels. The result proved in Theorem 12.1.1 is that the mutual information is additive for any two classical channels, so that classical correlations cannot enhance it.

information of the classical tandem channel $\mathcal{N} \otimes \mathcal{N}$ is strictly greater than two individual mutual informations:

$$I(\mathcal{N} \otimes \mathcal{N}) \stackrel{?}{>} 2I(\mathcal{N}). \tag{12.8}$$

Figure 12.1 displays the scenario corresponding to the above question.

In fact, we can take the above argument to its extreme, by defining the regularized mutual information $I_{\text{reg}}(\mathcal{N})$ of a classical channel as follows:

$$I_{\text{reg}}(\mathcal{N}) \equiv \lim_{n\to\infty} \frac{1}{n} I\left(\mathcal{N}^{\otimes n}\right). \tag{12.9}$$

In the above definition, the quantity $I(\mathcal{N}^{\otimes n})$ is as follows:

$$I\left(\mathcal{N}^{\otimes n}\right) \equiv \max_{p_{X^n}(x^n)} I(X^n; Y^n), \tag{12.10}$$

$\mathcal{N}^{\otimes n}$ denotes n channels in tandem with mapping

$$p_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} p_{Y_i|X_i}(y_i|x_i), \tag{12.11}$$

where $X^n \equiv X_1, X_2, \ldots, X_n$, $x^n \equiv x_1, x_2, \ldots, x_n$, and $Y^n \equiv Y_1, Y_2, \ldots, Y_n$. The potential superadditive effect would have the following form after bootstrapping the inequality in (12.8) to the regularization:

$$I_{\text{reg}}(\mathcal{N}) \stackrel{?}{>} I(\mathcal{N}). \tag{12.12}$$

Exercise 12.1.1 Determine the maximum value of $I_{\text{reg}}(\mathcal{N})$ when taking the limit. Thus, this quantity is finite.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The next section shows that the above strict inequalities do not hold for a classical channel, implying that no such superadditive effect occurs for its mutual information. In fact, the mutual information of a classical channel obeys an additivity property that is the cornerstone of our understanding of classical information theory. This additivity property implies that

$$I(\mathcal{N} \otimes \mathcal{N}) = 2I(\mathcal{N}) \tag{12.13}$$

and

$$I_{\text{reg}}(\mathcal{N}) = I(\mathcal{N}), \tag{12.14}$$

by an inductive argument. Thus, classical correlations between inputs do not increase the mutual information of a classical channel.

We are stressing the importance of additivity in classical information theory because recent research has demonstrated that superadditive effects can occur in quantum Shannon theory (see Section 19.5, for example). These quantum results imply that our understanding of quantum Shannon theory is not yet complete, but they also demonstrate the fascinating possibility that quantum correlations can increase the information throughput of a quantum channel.

12.1.2 Additivity

The mutual information of classical channels satisfies the important and natural property of additivity. We prove the strongest form of additivity that occurs for the mutual information of two different classical channels. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ denote two *different* classical channels corresponding to the respective mappings $p_{Y_1|X_1}(y_1|x_1)$ and $p_{Y_2|X_2}(y_2|x_2)$, and let $\mathcal{N}_1 \otimes \mathcal{N}_2$ denote the tandem channel that corresponds to the mapping

$$p_{Y_1,Y_2|X_1,X_2}(y_1, y_2|x_1, x_2) = p_{Y_1|X_1}(y_1|x_1)p_{Y_2|X_2}(y_2|x_2). \tag{12.15}$$

The mutual information of the tandem channel is then as follows:

$$I(\mathcal{N}_1 \otimes \mathcal{N}_2) \equiv \max_{p_{X_1,X_2}(x_1,x_2)} I(X_1, X_2; Y_1, Y_2). \tag{12.16}$$

The following theorem states the additivity property.

Theorem 12.1.1 (Additivity of Mutual Information of Classical Channels)**.** *The mutual information of the classical tandem channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is the sum of their individual mutual informations:*

$$I(\mathcal{N}_1 \otimes \mathcal{N}_2) = I(\mathcal{N}_1) + I(\mathcal{N}_2). \tag{12.17}$$

Proof. We first prove the inequality $I(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq I(\mathcal{N}_1) + I(\mathcal{N}_2)$. This inequality is more trivial to prove than the other direction. Let $p^*_{X_1}(x_1)$ and $p^*_{X_2}(x_2)$ denote the distributions that achieve the respective maximums of $I(\mathcal{N}_1)$ and $I(\mathcal{N}_2)$. The joint probability distribution for all input and output random variables is then as follows:

$$p_{X_1,X_2,Y_1,Y_2}(x_1, x_2, y_1, y_2) = p^*_{X_1}(x_1)p^*_{X_2}(x_2)p_{Y_1|X_1}(y_1|x_1)p_{Y_2|X_2}(y_2|x_2). \tag{12.18}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Observe that X_1 and Y_1 are independent of X_2 and Y_2. Then the following chain of inequalities holds:

$$\begin{aligned} I(\mathcal{N}_1) + I(\mathcal{N}_2) &= I(X_1;Y_1) + I(X_2;Y_2) && (12.19)\\ &= H(Y_1) - H(Y_1|X_1) + H(Y_2) - H(Y_2|X_2) && (12.20)\\ &= H(Y_1,Y_2) - H(Y_1|X_1,X_2) - H(Y_2|X_2,X_1,Y_1) && (12.21)\\ &= H(Y_1,Y_2) - H(Y_1,Y_2|X_1,X_2) && (12.22)\\ &= I(X_1,X_2;Y_1;Y_2) && (12.23)\\ &\leq I(\mathcal{N}_1 \otimes \mathcal{N}_2). && (12.24) \end{aligned}$$

The first equality follows by evaluating the mutual informations $I(\mathcal{N}_1)$ and $I(\mathcal{N}_2)$ with respect to the maximizing distributions $p^*_{X_1}(x_1)$ and $p^*_{X_2}(x_2)$. The second equality follows by expanding the mutual informations. The third equality follows because $H(Y_1)+H(Y_2) = H(Y_1,Y_2)$ when random variables Y_1 and Y_2 are independent, $H(Y_1|X_1,X_2) = H(Y_1|X_1)$ when Y_1 is independent of X_2, and $H(Y_2|X_2,X_1,Y_1) = H(Y_2|X_2)$ when Y_2 is independent of X_1 and Y_1. The fourth equality follows from the entropy chain rule in Exercise 10.3.2. The next equality follows again by expanding the mutual information. The final inequality follows because the input distribution $p^*_{X_1}(x_1)p^*_{X_2}(x_2)$ is a particular input distribution of the more general form $p_{X_1,X_2}(x_1,x_2)$ needed in the maximization of the mutual information of the tandem channel $\mathcal{N}_1 \otimes \mathcal{N}_2$. We now prove the non-trivial inequality $I(\mathcal{N}_1 \otimes \mathcal{N}_2) \leq I(\mathcal{N}_1) + I(\mathcal{N}_2)$. Let $p^*_{X_1,X_2}(x_1,x_2)$ denote the distribution that maximizes $I(\mathcal{N}_1 \otimes \mathcal{N}_2)$, and let

$$q_{X_1|X_2}(x_1|x_2) \quad \text{and} \quad q_{X_2}(x_2) \tag{12.25}$$

be distributions such that

$$p^*_{X_1,X_2}(x_1,x_2) = q_{X_1|X_2}(x_1|x_2)q_{X_2}(x_2). \tag{12.26}$$

Recall that the mapping for the tandem channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is as follows:

$$p_{Y_1,Y_2|X_1,X_2}(y_1,y_2|x_1,x_2) = p_{Y_1|X_1}(y_1|x_1)p_{Y_2|X_2}(y_2|x_2). \tag{12.27}$$

By summing over y_2, we observe that Y_1 and X_2 are independent because

$$p_{Y_1|X_1,X_2}(y_1|x_1,x_2) = p_{Y_1|X_1}(y_1|x_1). \tag{12.28}$$

Also, the joint distribution $p_{X_1,Y_1,Y_2|X_2}(x_1,y_1,y_2|x_2)$ has the form

$$p_{X_1,Y_1,Y_2|X_2}(x_1,y_1,y_2|x_2) = p_{Y_1|X_1}(y_1|x_1)q_{X_1|X_2}(x_1|x_2)p_{Y_2|X_2}(y_2|x_2). \tag{12.29}$$

Then Y_2 is conditionally independent of X_1 and Y_1 when conditioning on X_2. Consider the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

following chain of inequalities:

$$\begin{aligned} I(\mathcal{N}_1 \otimes \mathcal{N}_2) &= I(X_1, X_2; Y_1; Y_2) && (12.30)\\ &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) && (12.31)\\ &= H(Y_1, Y_2) - H(Y_1|X_1, X_2) - H(Y_2|Y_1, X_1, X_2) && (12.32)\\ &= H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2) && (12.33)\\ &\leq H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) && (12.34)\\ &= I(X_1; Y_1) + I(X_2; Y_2) && (12.35)\\ &\leq I(\mathcal{N}_1) + I(\mathcal{N}_2). && (12.36) \end{aligned}$$

The first equality follows from the definition of $I(\mathcal{N}_1 \otimes \mathcal{N}_2)$ in (12.16) and by evaluating the mutual information with respect to the distributions $p^*_{X_1,X_2}(x_1, x_2)$, $p_{Y_1|X_1}(y_1|x_1)$, and $p_{Y_2|X_2}(y_2|x_2)$. The second equality follows by expanding the mutual information $I(X_1, X_2; Y_1; Y_2)$. The third equality follows from the entropy chain rule. The fourth equality follows because $H(Y_1|X_1, X_2) = H(Y_1|X_1)$ when Y_1 is independent of X_2 as pointed out in (12.27). Also, the equality follows because $H(Y_2|Y_1, X_1, X_2) = H(Y_2|X_2)$ when Y_2 is conditionally independent of X_1 and Y_1 as pointed out in (12.29). The first inequality follows from subadditivity of entropy (Exercise 10.3.3). The last equality follows from the definition of mutual information, and the final inequality follows because the marginal distributions for X_1 and X_2 can only achieve a mutual information less than the respective maximizing marginal distributions for $I(\mathcal{N}_1)$ and $I(\mathcal{N}_2)$. □

A simple corollary of Theorem 12.1.1 is that correlations between input random variables cannot increase the mutual information of a classical channel. The proof follows by a straightforward induction argument. Thus, the *single-letter* expression in (12.5) for the mutual information of a classical channel suffices for understanding the ability of a classical channel to maintain correlations between its input and output.

Corollary 12.1.1. *The regularized mutual information of a classical channel is equal to its mutual information:*

$$I_{reg}(\mathcal{N}) = I(\mathcal{N}). \tag{12.37}$$

Proof. We prove the result using induction on n, by showing that $I(\mathcal{N}^{\otimes n}) = nI(\mathcal{N})$ for all n, implying that the limit in (12.9) is not necessary. The base case for $n = 1$ is trivial. Suppose the result holds for n: $I(\mathcal{N}^{\otimes n}) = nI(\mathcal{N})$. The following chain of equalities then proves the inductive step:

$$\begin{aligned} I(\mathcal{N}^{\otimes n+1}) &= I(\mathcal{N} \otimes \mathcal{N}^{\otimes n}) && (12.38)\\ &= I(\mathcal{N}) + I(\mathcal{N}^{\otimes n}) && (12.39)\\ &= I(\mathcal{N}) + nI(\mathcal{N}). && (12.40) \end{aligned}$$

The first equality follows because the channel $\mathcal{N}^{\otimes n+1}$ is equivalent to a tandem of $\mathcal{N}$ and $\mathcal{N}^{\otimes n}$. The second critical equality follows from the application of Theorem 12.1.1 because the distributions of $\mathcal{N}$ and $\mathcal{N}^{\otimes n}$ factorize as in (12.27). The final equality follows from the induction hypothesis. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

12.1.3 Optimizing the Mutual Information of a Classical Channel

The definition in (12.5) seems like a suitable definition for the mutual information of a classical channel, but how difficult is the maximization problem that it sets out? Theorem 12.1.2 below states an important property of the mutual information $I(X;Y)$ that allows us to answer this question. Suppose that we fix the conditional density $p_{Y|X}(y|x)$, but can vary the input density $p_X(x)$—this scenario is the same as the above one. Theorem 12.1.2 below proves that the mutual information $I(X;Y)$ is a concave function of the density $p_X(x)$. In particular, this result implies that the channel mutual information $I(\mathcal{N})$ has a unique global maximum, and the optimization problem is therefore a straightforward computation that can exploit convex optimization methods.

Theorem 12.1.2. *Suppose that we fix the conditional probability density $p_{Y|X}(y|x)$. Then the mutual information $I(X;Y)$ is concave in the marginal density $p_X(x)$:*

$$\lambda I(X_1;Y) + (1-\lambda) I(X_2;Y) \leq I(Z;Y), \tag{12.41}$$

where random variable X_1 has density $p_{X_1}(x)$, X_2 has density $p_{X_2}(x)$, and Z has density $\lambda p_{X_1}(x) + (1-\lambda) p_{X_2}(x)$.

Proof. Let us fix the density $p_{Y|X}(y|x)$. The density $p_Y(y)$ is a linear function of $p_X(x)$ because $p_Y(y) = \sum_x p_{Y|X}(y|x) p_X(x)$. Thus $H(Y)$ is concave in $p_X(x)$. Recall that the conditional entropy $H(Y|X) = \sum_x p_X(x) H(Y|X=x)$. The entropy $H(Y|X=x)$ is fixed when the conditional probability density $p_{Y|X}(y|x)$ is fixed. Thus, $H(Y|X)$ is a linear function of $p_X(x)$. These two results imply that the mutual information $I(X;Y)$ is concave in the marginal density $p_X(x)$ when the conditional density $p_{Y|X}(y|x)$ is fixed. □

12.2 Private Information of a Wiretap Channel

Suppose now that we extend the above two-user classical communication scenario to a three-user communication scenario, where the parties are Alice, Bob, and Eve. Alice would like to communicate to Bob while keeping her messages private from Eve. The channel $\mathcal{N}$ in this setting is the wiretap channel, corresponding to the following conditional probability density:

$$p_{Y,Z|X}(y,z|x). \tag{12.42}$$

Alice has access to the input random variable X, Bob receives output random variable Y, and Eve receives the random variable Z. Figure 12.2 depicts this setting.

We would like to establish a measure of information throughput for this scenario. It might seem intuitive that it should be the amount of correlations that Alice can establish with Bob, less the correlations that Eve receives:

$$I(X;Y) - I(X;Z). \tag{12.43}$$

But Alice can maximize over all possible coding strategies on her end (all possible probability distributions $p_X(x)$). This leads us to the following definition:

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

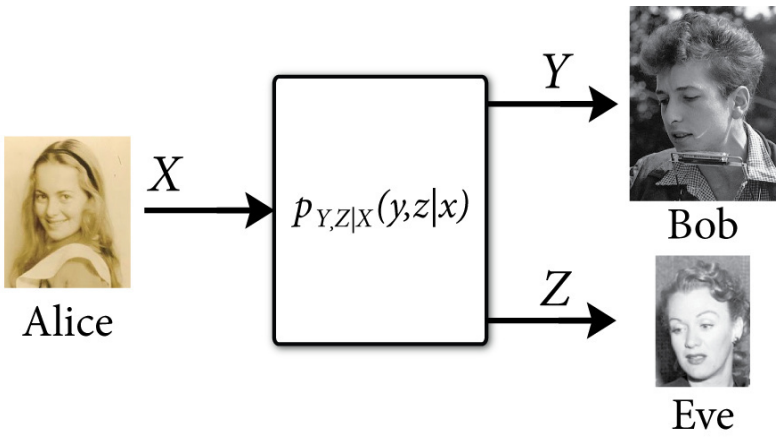

Figure 12.2: The setting for the classical wiretap channel.

Definition 12.2.1 (Private Information of a Wiretap Channel)**.** *The private information $P(\mathcal{N})$ of a classical wiretap channel is as follows:*

$$P(\mathcal{N}) \equiv \max_{p_X(x)} I(X;Y) - I(X;Z). \tag{12.44}$$

We should note that the above definition of the private information is not the most general formula—we could include a preprocessing step with the Markov chain $U \to X \to (Y,Z)$, but we stick with the above definition for simplicity.

It is possible to provide an operational interpretation of the private information, showing that it is indeed the private capacity of the wiretap channel, but we do not do that here. We instead focus on the additivity properties of the private information $P(\mathcal{N})$.

One may wonder if the above quantity is positive, given that it is the difference of two mutual informations. Positivity does hold, and a simple proof demonstrates this fact.

Property 12.2.1 The private information $P(\mathcal{N})$ of a wiretap channel is positive:

$$P(\mathcal{N}) \geq 0. \tag{12.45}$$

Proof. We can choose the density $p_X(x)$ in the maximization of $P(\mathcal{N})$ to be the degenerate distribution $p_X(x) = \delta_{x,x_0}$ for some realization x_0. Then both mutual informations $I(X;Y)$ and $I(X;Z)$ vanish, and their difference vanishes as well. The private information $P(\mathcal{N})$ can only then be greater than or equal to zero because the above choice $p_X(x)$ is a particular choice of the density $p_X(x)$ and $P(\mathcal{N})$ requires a maximization over all such distributions. □

12.2.1 Additivity of Private Information for Degraded Wiretap Channels

It is difficult to show that the private information of general wiretap channels is additive, but it is straightforward to do so for a particular type of wiretap channel, called a physically

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

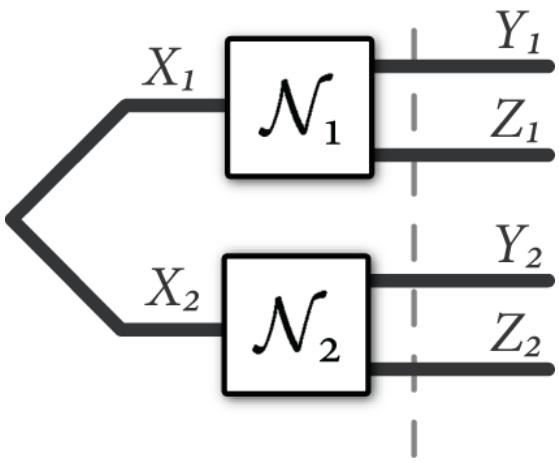

Figure 12.3: The above figure displays the scenario for determining whether the private information of two classical channels $\mathcal{N}_1$ and $\mathcal{N}_2$ is additive. The question of additivity is equivalent to the possibility of classical correlations being able to enhance the private information of two classical channels. The result proved in Theorem 12.2.1 is that the private information is additive for two degraded wiretap channels, so that classical correlations cannot enhance the private information in this case.

degradable wiretap channel. A wiretap channel is physically degradable if X, Y, and Z form the following Markov chain: $X \to Y \to Z$. That is, there is some channel $p_{Z|Y}(z|y)$ that Bob can apply to his output to simulate the channel $p_{Z|X}(z|x)$ to Eve:

$$p_{Z|X}(z|x) = p_{Z|Y}(z|y)p_{Y|X}(y|x). \tag{12.46}$$

This condition allows us to apply the data processing inequality to demonstrate that the private information of degraded wiretap channels is additive. Figure 12.3 displays the scenario corresponding to the analysis involved in determining whether the private information is additive.

Theorem 12.2.1 (Additivity of Private Information of Degraded Wiretap Channels)**.** *The private information of the classical tandem channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is the sum of their individual private informations:*

$$P(\mathcal{N}_1 \otimes \mathcal{N}_2) = P(\mathcal{N}_1) + P(\mathcal{N}_2). \tag{12.47}$$

Proof. The inequality $P(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq P(\mathcal{N}_1) + P(\mathcal{N}_2)$ is trivial and we leave it as an exercise for the reader to complete. We thus prove the non-trivial inequality for the case of degraded wiretap channels: $P(\mathcal{N}_1 \otimes \mathcal{N}_2) \leq P(\mathcal{N}_1) + P(\mathcal{N}_2)$. Let $p^*_{X_1,X_2}(x_1, x_2)$ be the distribution that maximizes the quantity $P(\mathcal{N}_1 \otimes \mathcal{N}_2)$. The channels are of the following form:

$$\begin{aligned} &p_{Y_1,Z_1|X_1}(y_1, z_1|x_1)p_{Y_2,Z_2|X_2}(y_2, z_2|x_2) \\ &\qquad = p_{Z_1|Y_1}(z_1|y_1)p_{Y_1|X_1}(y_1|x_1)p_{Z_2|Y_2}(z_2|y_2)p_{Y_2|X_2}(y_2|x_2). \end{aligned} \tag{12.48}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Observe then that Y_1 and Z_1 are independent of X_2. Also, Y_2 and Z_2 are independent of Y_1, Z_1, and X_1. Then the following chain of inequalities holds:

$$\begin{aligned}
&P(\mathcal{N}_1 \otimes \mathcal{N}_2) \\
&= I(X_1X_2; Y_1Y_2) - I(X_1X_2; Z_1Z_2) && (12.49)\\
&= H(Y_1Y_2) - H(Y_1Y_2|X_1X_2) - H(Z_1Z_2) + H(Z_1Z_2|X_1X_2) && (12.50)\\
&= H(Y_1Y_2) - H(Y_1|X_1X_2) - H(Y_2|Y_1X_1X_2) \\
&\quad - H(Z_1Z_2) + H(Z_1|X_1X_2) + H(Z_2|Z_1X_1X_2) && (12.51)\\
&= H(Y_1Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\
&\quad - H(Z_1Z_2) + H(Z_1|X_1) + H(Z_2|X_2) && (12.52)\\
&= H(Y_1) + H(Y_2) - I(Y_1; Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\
&\quad - H(Z_1) - H(Z_2) + I(Z_1; Z_2) + H(Z_1|X_1) + H(Z_2|X_2) && (12.53)\\
&\leq H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\
&\quad - H(Z_1) - H(Z_2) + H(Z_1|X_1) + H(Z_2|X_2) && (12.54)\\
&= I(X_1; Y_1) - I(X_1; Z_1) + I(X_2; Y_2) - I(X_2; Z_2) && (12.55)\\
&\leq P(\mathcal{N}_1) + P(\mathcal{N}_2) && (12.56)
\end{aligned}$$

The first equality follows from evaluating $P(\mathcal{N}_1 \otimes \mathcal{N}_2)$ on the maximizing distribution $p^*_{X_1,X_2}(x_1, x_2)$. The second equality follows by expanding the mutual informations. The third equality follows from the entropy chain rule, and the fourth follows because Y_1 and Z_1 are independent of X_2 and Y_2 and Z_2 are independent of Y_1, Z_1, and X_1. The fifth equality follows by replacing the joint entropies with the sum of the marginal entropies reduced by the mutual information. The important inequality follows because there is a degrading map from Y_1 to Z_1 and from Y_2 to Z_2, implying that $I(Y_1; Y_2) \geq I(Z_1; Z_2)$ by the data processing inequality. The last equality follows by combining the entropies into mutual informations, and the final inequality follows because these information quantities must be less than the same information quantities evaluated with respect to the maximizing distributions. □

An analogous notion of degradability exists in the quantum setting, and Section 12.5 demonstrates that degradable quantum channels have additive coherent information. The coherent information of a quantum channel is a measure of how much quantum information a sender can transmit through that channel to a receiver and thus is an important quantity to consider for quantum data transmission.

Exercise 12.2.1 Show that the sum of the individual private informations can never be greater than the private information of the classical tandem channel:

$$P(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq P(\mathcal{N}_1) + P(\mathcal{N}_2). \tag{12.57}$$

12.3 Holevo Information of a Quantum Channel

We now turn our attention to the case of dynamic informational measures for quantum channels, and we begin with a measure of classical correlations. Suppose that Alice would

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

like to establish classical correlations with Bob, and she wishes to exploit a quantum channel to do so. Alice can prepare an ensemble $\{p_X(x), \rho_x\}$ in her laboratory, where the states ρ_x are acceptable inputs to the quantum channel. She keeps a copy of the classical index x in some classical register X. The expected density operator of this ensemble is the following classical-quantum state:

$$\rho^{XA'} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^{A'}. \tag{12.58}$$

Such a preparation is the most general way that Alice can correlate classical data with a quantum state to input to the channel. Let ρ^{XB} be the state that arises from sending the A' system through the quantum channel $\mathcal{N}^{A'\to B}$:

$$\rho^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A'\to B}(\rho_x^{A'}). \tag{12.59}$$

We would like to determine a measure of the ability of the quantum channel to preserve classical correlations. We can appeal to ideas from the classical case in Section 12.1, while incorporating the static quantum measures from Chapter 11. A good measure of the input-output classical correlations is the Holevo information of the above classical quantum state: $I(X;B)_\rho$. This measure corresponds to a particular preparation that Alice chooses, but observe that she can prepare the input ensemble in such a way as to achieve the highest possible correlations. Maximizing the Holevo information over all possible preparations gives a measure called the Holevo information of the channel.

Definition 12.3.1 (Holevo Information of a Quantum Channel)**.** *The Holevo information of the channel is a measure of the classical correlations that Alice can establish with Bob:*

$$\chi(\mathcal{N}) \equiv \max_{\rho^{XA'}} I(X;B)_\rho, \tag{12.60}$$

where the maximization is over all input ensembles.

12.3.1 Additivity of the Holevo information for Specific Channels

The Holevo information of a quantum channel is generally not additive. The question of additivity for this case is *not* whether classical correlations can enhance the Holevo information, but it *is rather* whether quantum correlations can enhance it. That is, Alice can choose an ensemble of the form $\{p_X(x), \rho_x^{A_1'A_2'}\}$ for input to two uses of the quantum channel. The conditional density operators $\rho_x^{A_1'A_2'}$ can be entangled and these quantum correlations can potentially increase the Holevo information.

The question of additivity of the Holevo information of a quantum channel was a long-standing open conjecture in quantum information theory—many researchers thought that quantum correlations would not enhance it and that additivity would hold. But recent research has demonstrated a counterexample to the additivity conjecture, and perhaps unsurprisingly in hindsight, this counterexample exploits maximally entangled states to demonstrate superadditivity (see Section 19.5). Figure 12.4 displays the scenario corresponding to the question of additivity of the Holevo information.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

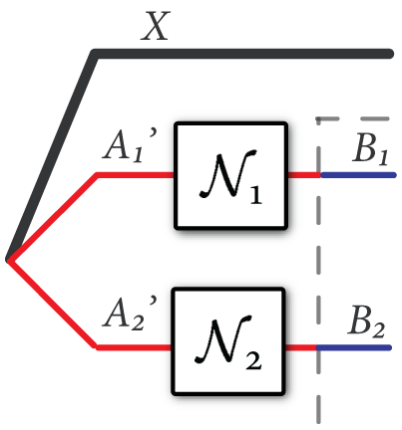

Figure 12.4: The above figure displays the scenario for determining whether the Holevo information of two quantum channels $\mathcal{N}_1$ and $\mathcal{N}_2$ is additive. The question of additivity is equivalent to the possibility of quantum correlations being able to enhance the Holevo information of two quantum channels. The result proved in Theorem 12.3.1 is that the Holevo information is additive for the tensor product of an entanglement-breaking channel and any other quantum channel, so that quantum correlations cannot enhance the Holevo information in this case. This is perhaps intuitive because an entanglement breaking channel destroys quantum correlations in the form of quantum entanglement.

Additivity of Holevo information may not hold for all quantum channels, but it is possible to prove its additivity for certain classes of quantum channels. One such class for which additivity holds is the class of entanglement-breaking channels, and the proof of additivity is perhaps the simplest for this case.

Theorem 12.3.1 (Additivity of the Holevo information of an EB Channel)**.** *Suppose that a quantum channel $\mathcal{N}_1$ is entanglement-breaking. Then the Holevo information $\chi(\mathcal{N}_1 \otimes \mathcal{N}_2)$ of the tensor product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is the sum of the individual Holevo informations $\chi(\mathcal{N}_1)$ and $\chi(\mathcal{N}_2)$:*

$$\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) = \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2). \tag{12.61}$$

Proof. The trivial inequality $\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2)$ holds for any two quantum channels $\mathcal{N}_1$ and $\mathcal{N}_2$ because we can choose the input ensemble on the LHS to be a tensor product of the ones that individually maximize the terms on the RHS. We now prove the nontrivial inequality $\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) \leq \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2)$ that holds when $\mathcal{N}_1$ is entanglement-breaking. Let $\rho^{XB_1B_2}$ be the state that maximizes the Holevo information $\chi(\mathcal{N}_1 \otimes \mathcal{N}_2)$, where

$$\rho^{XB_1B_2} \equiv (\mathcal{N}_1^{A_1' \to B_1} \otimes \mathcal{N}_2^{A_2' \to B_2})(\rho^{XA_1'A_2'}), \tag{12.62}$$

$$\rho^{XA_1'A_2'} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^{A_1'A_2'}. \tag{12.63}$$

The action of $\mathcal{N}_1$ is to break entanglement. Let $\rho^{XB_1A_2'}$ be the state after only the entanglement-

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

breaking channel $\mathcal{N}_1$ acts. We can write this state as follows:

$$\begin{aligned}
\rho^{XB_1A_2'} &\equiv \mathcal{N}_1^{A_1'\to B_1}(\rho^{XA_1'A_2'}) && (12.64)\\
&= \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}_1^{A_1'\to B_1}(\rho_x^{A_1'A_2'}) && (12.65)\\
&= \sum_x p_X(x)|x\rangle\langle x|^X \otimes \sum_y p_{Y|X}(y|x)\ \sigma_{x,y}^{B_1}\otimes\theta_{x,y}^{A_2'} && (12.66)\\
&= \sum_{x,y} p_{Y|X}(y|x)p_X(x)|x\rangle\langle x|^X \otimes \sigma_{x,y}^{B_1}\otimes\theta_{x,y}^{A_2'}. && (12.67)
\end{aligned}$$

The third equality follows because the channel $\mathcal{N}_1$ breaks any entanglement in the state $\rho_x^{A_1'A_2'}$, leaving behind a separable state $\sum_y p_{Y|X}(y|x)\ \sigma_{x,y}^{B_1}\otimes\theta_{x,y}^{A_2'}$. Then the state $\rho^{XB_1B_2}$ has the form

$$\rho^{XB_1B_2} = \sum_{x,y} p_{Y|X}(y|x)p_X(x)|x\rangle\langle x|^X \otimes \sigma_{x,y}^{B_1}\otimes\mathcal{N}_2^{A_2'\to B_2}(\theta_{x,y}^{A_2'}). \tag{12.68}$$

Let $\omega^{XYB_1B_2}$ be an extension of $\rho^{XB_1B_2}$ where

$$\omega^{XYB_1B_2} \equiv \sum_{x,y} p_{Y|X}(y|x)p_X(x)|x\rangle\langle x|^X \otimes |y\rangle\langle y|^Y \otimes \sigma_{x,y}^{B_1}\otimes\mathcal{N}_2^{A_2'\to B_2}(\theta_{x,y}^{A_2'}), \tag{12.69}$$

and $\mathrm{Tr}_Y\{\omega^{XYB_1B_2}\} = \rho^{XB_1B_2}$. Then the following chain of inequalities holds:

$$\begin{aligned}
\chi(\mathcal{N}_1\otimes\mathcal{N}_2) &= I(X;B_1B_2)_\rho && (12.70)\\
&= I(X;B_1B_2)_\omega && (12.71)\\
&\leq I(XY;B_1B_2)_\omega && (12.72)\\
&= H(B_1B_2)_\omega - H(B_1B_2|XY)_\omega && (12.73)\\
&= H(B_1B_2)_\omega - H(B_1|XY)_\omega - H(B_2|XY)_\omega && (12.74)\\
&\leq H(B_1)_\omega + H(B_2)_\omega - H(B_1|XY)_\omega - H(B_2|XY)_\omega && (12.75)\\
&= I(XY;B_1)_\omega + I(XY;B_2)_\omega && (12.76)\\
&\leq \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2). && (12.77)
\end{aligned}$$

The first equality follows because $\rho^{XB_1B_2}$ is the state that maximizes the Holevo information $\chi(\mathcal{N}_1\otimes\mathcal{N}_2)$ of the tensor product channel $\mathcal{N}_1\otimes\mathcal{N}_2$. The second equality follows because the reduced state of $\omega^{XYB_1B_2}$ on systems X, B_1, and B_2 is equal to $\rho^{XB_1B_2}$. The first inequality follows from the quantum data processing inequality. The third equality follows by expanding the mutual information $I(XY;B_1B_2)_\omega$. The fourth equality is the crucial one that exploits the entanglement-breaking property. It follows by examining (12.68) and observing that the state $\omega^{XYB_1B_2}$ on systems B_1 and B_2 is product when conditioned on classical variables X and Y. The second inequality follows from subadditivity of entropy. The last equality follows from straightforward entropic manipulations, and the final inequality follows because ω^{XYB_1} is a particular state of the form needed in the maximization of $\chi(\mathcal{N}_1)$, and the same holds for the state ω^{XYB_2}. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Corollary 12.3.1. *The regularized Holevo information of an entanglement-breaking quantum channel $\mathcal{N}$ is equal to its Holevo information:*

$$\chi_{reg}(\mathcal{N}) = \chi(\mathcal{N}). \tag{12.78}$$

Proof. The proof of this property uses the same induction argument as in Corollary 12.1.1 and exploits the additivity property in Theorem 12.3.1 above. □

12.3.2 Optimizing the Holevo Information

Pure States are Sufficient

The following theorem allows us to simplify the optimization problem that (12.60) sets out—we show that it is sufficient to consider ensembles of pure states at the input.

Theorem 12.3.2. *The Holevo information is equivalent to a maximization over only pure states:*

$$\chi(\mathcal{N}) = \max_{\rho^{XA'}} I(X;B) = \max_{\tau^{XA'}} I(X;B), \tag{12.79}$$

where

$$\tau^{XA'} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes |\phi_x\rangle\langle\phi_x|^{A'}, \tag{12.80}$$

and τ^{XB} is the state that results from sending the A' system of the above state through the quantum channel $\mathcal{N}^{A'\to B}$.

Proof. Suppose that $\rho^{XA'}$ is any state of the form in (12.58). Consider a spectral decomposition of the states $\rho_x^{A'}$:

$$\rho_x^{A'} = \sum_y p_{Y|X}(y|x)\psi_{x,y}^{A'}, \tag{12.81}$$

where the states $\psi_{x,y}^{A'}$ are pure. Then let $\sigma^{XYA'}$ denote the following state:

$$\sigma^{XYA'} \equiv \sum_x p_{Y|X}(y|x)p_X(x)|x\rangle\langle x|^X \otimes |y\rangle\langle y|^Y \otimes \psi_{x,y}^{A'}, \tag{12.82}$$

so that $\mathrm{Tr}_Y\{\sigma^{XYA'}\} = \rho^{XA'}$. Also, observe that $\sigma^{XYA'}$ is a state of the form $\tau^{XA'}$ with XY as the classical system. Let σ^{XYB} denote the state that results from sending the A' system through the quantum channel $\mathcal{N}^{A'\to B}$. Then the following relations hold

$$I(X;B)_\rho = I(X;B)_\sigma \tag{12.83}$$

$$\leq I(XY;B)_\sigma \tag{12.84}$$

The equality follows because $\mathrm{Tr}_Y\{\sigma^{XYB}\} = \rho^{XB}$ and the inequality follows from the quantum data processing inequality. It then suffices to consider ensembles with only pure states because the state σ^{XYB} is a state of the form τ^{XB} with the combined system XY acting as the classical system. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Concavity in the Distribution and Convexity in the Signal States

We now show that the Holevo information is concave as a function of the input distribution when the signal states are fixed.

Theorem 12.3.3. *The Holevo information $I(X;B)$ is concave in the input distribution when the signal states are fixed, in the sense that*

$$\lambda I(X;B)_{\sigma_0} + (1-\lambda) I(X;B)_{\sigma_1} \leq I(X;B)_{\sigma}, \tag{12.85}$$

where σ_0^{XB} and σ_1^{XB} are of the form

$$\sigma_0^{XB} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes \mathcal{N}(\sigma_x), \tag{12.86}$$

$$\sigma_1^{XB} \equiv \sum_x q_X(x) |x\rangle\langle x|^X \otimes \mathcal{N}(\sigma_x), \tag{12.87}$$

and σ^{XB} is a mixture of the states σ_0^{XB} and σ_1^{XB} of the form:

$$\sigma^{XB} \equiv \sum_x [\lambda p_X(x) + (1-\lambda) q_X(x)] |x\rangle\langle x|^X \otimes \mathcal{N}(\sigma_x), \tag{12.88}$$

where $0 \leq \lambda \leq 1$.

Proof. Let σ^{XUB} be the state

$$\sigma^{UXB} \equiv \sum_x \Big[p_X(x) |x\rangle\langle x|^X \otimes \lambda |0\rangle\langle 0|^U + q_X(x) |x\rangle\langle x|^X \otimes (1-\lambda) |1\rangle\langle 1|^U \Big] \otimes \mathcal{N}(\sigma_x). \tag{12.89}$$

Observe that $\text{Tr}_U\{\sigma^{XUB}\} = \sigma^{XB}$. Then the statement of concavity is equivalent to

$$I(X;B|U)_{\sigma} \leq I(X;B)_{\sigma}. \tag{12.90}$$

We can rewrite this as

$$H(B|U)_{\sigma} - H(B|UX)_{\sigma} \leq H(B)_{\sigma} - H(B|X)_{\sigma}. \tag{12.91}$$

Observe that

$$H(B|UX)_{\sigma} = H(B|X)_{\sigma}, \tag{12.92}$$

i.e., one can calculate that both of these are equal to

$$\sum_x [\lambda p_X(x) + (1-\lambda) q_X(x)] H(\mathcal{N}(\sigma_x)). \tag{12.93}$$

The statement of concavity then becomes

$$H(B|U)_{\sigma} \leq H(B)_{\sigma}, \tag{12.94}$$

which follows from concavity of quantum entropy. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The Holevo information is convex as a function of the signal states when the input distribution is fixed.

Theorem 12.3.4. *The Holevo information $I(X;B)$ is convex in the signal states when the input distribution is fixed, in the sense that*

$$\lambda I(X;B)_{\sigma_0} + (1-\lambda) I(X;B)_{\sigma_1} \geq I(X;B)_{\sigma}, \tag{12.95}$$

where σ_0^{XB} and σ_1^{XB} are of the form

$$\sigma_0^{XB} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes \mathcal{N}(\sigma_x), \tag{12.96}$$

$$\sigma_1^{XB} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes \mathcal{N}(\omega_x), \tag{12.97}$$

and σ^{XB} is a mixture of the states σ_0^{XB} and σ_1^{XB} of the form:

$$\sigma^{XB} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes \mathcal{N}(\lambda\sigma_x + (1-\lambda)\omega_x), \tag{12.98}$$

where $0 \leq \lambda \leq 1$.

Proof. Let σ^{XUB} be the state

$$\sigma^{XUB} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes \Big[\lambda |0\rangle\langle 0|^U \otimes \mathcal{N}(\sigma_x) + (1-\lambda)|1\rangle\langle 1|^U \otimes \mathcal{N}(\omega_x)\Big]. \tag{12.99}$$

Observe that $\mathrm{Tr}_U\{\sigma^{XUB}\} = \sigma^{XB}$. Then convexity in the input states is equivalent to the statement

$$I(X;B|U)_{\sigma} \geq I(X;B)_{\sigma}. \tag{12.100}$$

Consider that

$$I(X;B|U)_{\sigma} = I(X;BU)_{\sigma} - I(X;U)_{\sigma}, \tag{12.101}$$

by the chain rule for the quantum mutual information. Since the input distribution $p_X(x)$ is fixed, there are no correlations between X and the convexity variable U, so that $I(X;U)_{\sigma} = 0$. Thus, the above inequality is equivalent to

$$I(X;BU)_{\sigma} \geq I(X;B)_{\sigma}, \tag{12.102}$$

which follows from the quantum data processing inequality. □

In the above two theorems, we have shown that the Holevo information is either concave or convex depending on whether the signal states or the input distribution are fixed, respectively. Thus, the computation of the Holevo information of a general quantum channel becomes difficult as the input dimension of the channel grows larger, since a local maximum of the Holevo information is not necessarily a global maximum. Though, if the channel has a classical input and a quantum output, the computation of the Holevo information is straightforward because the only input parameter is the input distribution, and we proved that the Holevo information is a concave function of the input distribution.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

12.4 Mutual Information of a Quantum Channel

We now consider a measure of the ability of a quantum channel to preserve quantum correlations. The way that we arrive at this measure is similar to what we have seen before. Alice prepares some pure quantum state $\phi^{AA'}$ in her laboratory, and inputs the A' system to a quantum channel $\mathcal{N}^{A'\to B}$—this transmission gives rise to the following noisy state:

$$\rho^{AB} = \mathcal{N}^{A'\to B}\left(\phi^{AA'}\right). \tag{12.103}$$

The quantum mutual information $I(A;B)_\rho$ is a static measure of quantum correlations present in the state ρ^{AB}. To maximize the quantum correlations that the quantum channel can establish, Alice should maximize the quantum mutual information $I(A;B)_\rho$ over all possible pure states that she can input to the channel $\mathcal{N}^{A'\to B}$. This procedure leads to the definition of the mutual information $I(\mathcal{N})$ of a quantum channel:

$$I(\mathcal{N}) \equiv \max_{\phi^{AA'}} I(A;B)_{\rho^{AB}} \tag{12.104}$$

The mutual information of a quantum channel corresponds to an important operational task that is not particularly obvious from the above discussion. Suppose that Alice and Bob share unlimited bipartite entanglement in whatever form they wish, and suppose they have access to a large number of independent uses of the channel $\mathcal{N}^{A'\to B}$. Then the mutual information of the channel corresponds to the maximal amount of classical information that they can transmit in such a setting. This setting is the noisy analog of the super-dense coding protocol from Chapter 6 (recall the discussion in Section 6.4). By teleportation, the maximal amount of quantum information that they can transmit is half of the mutual information of the channel. We discuss how to prove these statements rigorously in Chapter 20.

12.4.1 Additivity

There might be little reason to expect that the quantum mutual information of a quantum channel is additive, given that the Holevo information is not. But perhaps surprisingly, additivity does hold for the mutual information of a quantum channel! This result means that we completely understand this measure of information throughput, and it also means that we understand the operational task to which it corresponds (entanglement-assisted classical coding discussed in the previous section).

We might intuitively attempt to explain this phenomenon in terms of this operational task—Alice and Bob already share unlimited entanglement between their terminals and so entangled correlations at the input of the channel do not lead to any superadditive effect as it does for the Holevo information. This explanation is somewhat rough, but perhaps the additivity proof explains best why additivity holds. The crucial inequality in the proof follows from three applications of the strong subadditivity inequality (Theorem 11.9.1) and one application of subadditivity (Corollary 11.8.1). This highlights the importance of strong subadditivity in quantum Shannon theory. Figure 12.5 illustrates the setting corresponding to the analysis for additivity of the mutual information of a quantum channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

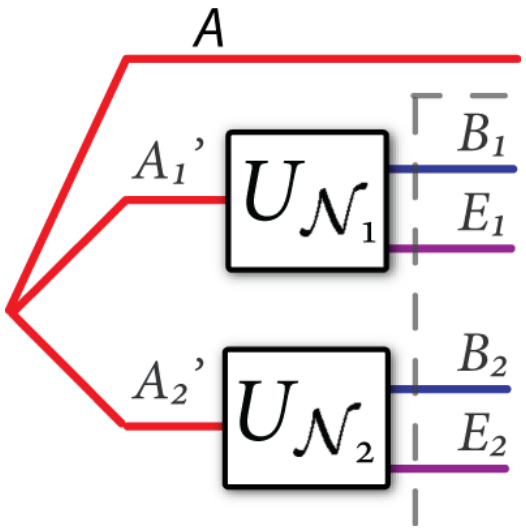

Figure 12.5: The above figure displays the scenario for determining whether the mutual information of two quantum channels $\mathcal{N}_1$ and $\mathcal{N}_2$ is additive. The question of additivity is equivalent to the possibility of quantum correlations between channel inputs being able to enhance the mutual information of two classical channels. The result proved in Theorem 12.4.1 is that the mutual information is additive for any two quantum channels, so that quantum correlations cannot enhance it.

Theorem 12.4.1 (Additivity of Quantum Mutual Information of Quantum Channels)**.** *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be any quantum channels. Then the mutual information of the tensor product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is the sum of their individual mutual informations:*

$$I(\mathcal{N}_1 \otimes \mathcal{N}_2) = I(\mathcal{N}_1) + I(\mathcal{N}_2). \tag{12.105}$$

Proof. We first prove the trivial inequality $I(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq I(\mathcal{N}_1) + I(\mathcal{N}_2)$. Let $\phi^{A_1 A_1'}$ and $\psi^{A_2 A_2'}$ be the states the maximize the respective mutual informations $I(\mathcal{N}_1)$ and $I(\mathcal{N}_2)$. Let $U_{\mathcal{N}_1}^{A_1' \to B_1 E_1}$ and $U_{\mathcal{N}_2}^{A_2' \to B_2 E_2}$ denote the respective isometric extensions of $\mathcal{N}_1$ and $\mathcal{N}_2$. The states ϕ and ψ then lead to a state φ where

$$\varphi^{A_1 A_2 B_1 B_2 E_1 E_2} \equiv (U_{\mathcal{N}_1}^{A_1' \to B_1 E_1} \otimes U_{\mathcal{N}_2}^{A_2' \to B_2 E_2})(\phi^{A_1 A_1'} \otimes \psi^{A_2 A_2'}). \tag{12.106}$$

Observe that the state $\text{Tr}_{E_1 E_2}\{\varphi\}$ is a particular state of the form required in the maximiza-

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

tion of $I(\mathcal{N}_1 \otimes \mathcal{N}_2)$, by taking $A \equiv A_1 A_2$. Then the following inequalities hold:

$$\begin{aligned}
I(\mathcal{N}_1) + I(\mathcal{N}_2) &= I(A_1; B_1)_{\mathcal{N}_1(\phi)} + I(A_2; B_2)_{\mathcal{N}_2(\psi)} && (12.107)\\
&= H(B_1)_{\mathcal{N}_1(\phi)} - H(B_1|A_1)_{\mathcal{N}_1(\phi)} \\
&\quad + H(B_2)_{\mathcal{N}_2(\psi)} - H(B_2|A_2)_{\mathcal{N}_2(\psi)} && (12.108)\\
&= H(B_1B_2)_{\varphi} - H(B_1|A_1A_2)_{\varphi} - H(B_2|A_2A_1B_1)_{\varphi} && (12.109)\\
&= H(B_1B_2)_{\varphi} - H(B_1B_1|A_1A_2)_{\varphi} && (12.110)\\
&= I(A_1A_2; B_1B_2)_{\varphi} && (12.111)\\
&\leq I(\mathcal{N}_1 \otimes \mathcal{N}_2). && (12.112)
\end{aligned}$$

The first equality follows by evaluating the mutual informations $I(\mathcal{N}_1)$ and $I(\mathcal{N}_2)$ with respect to the maximizing states $\phi^{A_1A_1'}$ and $\psi^{A_2A_2'}$. The second equality follows from the expansion of mutual information in (11.73). The third equality follows because $H(B_1) + H(B_2) = H(B_1B_2)$ when the quantum state φ on systems B_1 and B_2 is in a product state, $H(B_1|A_1A_2) = H(B_1|A_1)$ when the state φ on B_1 and A_1 is product with respect to A_2 (see Exercise 11.4.1), and $H(B_2|A_2A_1B_1)$ when the state φ on B_2 and A_2 is product with respect to A_1 and B_1 (again see Exercise 11.4.1). The fourth equality follows from the entropy chain rule. The fifth equality follows again from the expansion of mutual information. The final inequality follows because the input state $\phi^{A_1A_1'} \otimes \psi^{A_2A_2'}$ is a particular input state of the more general form $\phi^{AA_1'A_2'}$ needed in the maximization of the quantum mutual information of the tensor product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$. Notice that the steps here are almost identical to the ones in the classical proof of Theorem 12.1.1 with the exception that we use the quantum generalization of the classical properties. We now prove the non-trivial inequality $I(\mathcal{N}_1 \otimes \mathcal{N}_2) \leq I(\mathcal{N}_1) + I(\mathcal{N}_2)$. Let $\phi^{AA_1'A_2'}$ be the state that maximizes the mutual information $I(\mathcal{N}_1 \otimes \mathcal{N}_2)$ and let

$$\begin{aligned}
\sigma^{AB_1E_1A_2'} &\equiv U_{\mathcal{N}_1}^{A_1' \to B_1E_1}(\phi^{AA_1'A_2'}), && (12.113)\\
\theta^{AA_1'B_2E_2} &\equiv U_{\mathcal{N}_2}^{A_2' \to B_2E_2}(\phi^{AA_1'A_2'}), && (12.114)\\
\phi^{AB_1E_1B_2E_2} &\equiv (U_{\mathcal{N}_1}^{A_1' \to B_1E_1} \otimes U_{\mathcal{N}_2}^{A_2' \to B_2E_2})(\phi^{AA_1'A_2'}). && (12.115)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Consider the following chain of inequalities:

$$\begin{aligned}
I(\mathcal{N}_1 \otimes \mathcal{N}_2) &= I(A; B_1B_2)_\phi && (12.116)\\
&= H(A)_\phi + H(B_1B_2)_\phi - H(AB_1B_2)_\phi && (12.117)\\
&= H(B_1B_2E_1E_2)_\phi + H(B_1B_2)_\phi - H(E_1E_2)_\phi && (12.118)\\
&= H(B_1B_2|E_1E_2)_\phi + H(B_1B_2)_\phi && (12.119)\\
&\leq H(B_1|E_1)_\phi + H(B_2|E_2)_\phi + H(B_1)_\phi + H(B_2)_\phi && (12.120)\\
&= H(B_1E_1)_\phi + H(B_1)_\phi - H(E_1)_\phi \\
&\quad + H(B_2E_2)_\phi + H(B_2)_\phi - H(E_2)_\phi && (12.121)\\
&= H(AA_2')_\sigma + H(B_1)_\sigma - H(AA_2'B_1)_\sigma \\
&\quad + H(AA_1')_\theta + H(B_2)_\theta - H(AA_1'B_2)_\theta && (12.122)\\
&= I(AA_2'; B_1)_\sigma + I(AA_1'; B_2)_\theta && (12.123)\\
&\leq I(\mathcal{N}_1) + I(\mathcal{N}_2). && (12.124)
\end{aligned}$$

The first equality follows from the definition of $I(\mathcal{N}_1 \otimes \mathcal{N}_2)$ in (12.104) and evaluating $I(A; B_1B_2)$ with respect to the maximizing state ϕ. The second equality follows by expanding the quantum mutual information. The third equality follows because the state ϕ on systems A, B_1, B_2, E_1, and E_2 is pure. The fourth equality follows from the definition of conditional quantum entropy in (11.4.1). The first inequality is the crucial one that leads to additivity. It follows from three applications of strong subadditivity (Theorem 11.7.1) to obtain $H(B_1B_2|E_1E_2) \leq H(B_1|E_1) + H(B_2|E_2)$ and subadditivity (Theorem 11.6.1) to obtain $H(B_1B_2) \leq H(B_1) + H(B_2)$. The fifth equality follows by expanding the conditional quantum entropies $H(B_1|E_1)$ and $H(B_2|E_2)$. The sixth equality follows because the state θ on systems A, A_2', B_1, and E_1 is pure, and the state σ on systems A, A_1', B_2, and E_2 is pure. The last equality follows from the definition of quantum mutual information, and the final inequality follows because the states θ and σ are particular states of the form needed in the respective maximizations of $I(\mathcal{N}_1)$ and $I(\mathcal{N}_2)$. □

Corollary 12.4.1. *The regularized mutual information of any quantum channel is equal to its mutual information:*

$$I_{reg}(\mathcal{N}) = I(\mathcal{N}). \tag{12.125}$$

Proof. The proof of this property uses the same induction argument as in Corollary 12.1.1 and exploits the additivity property in Theorem 12.4.1 above. □

Exercise 12.4.1 (Alternate Mutual Information of a Quantum Channel) Let ρ^{XAB} denote a state of the following form:

$$\rho^{XAB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A'\to B}(\phi_x^{AA'}). \tag{12.126}$$

Consider the following alternate definition of the mutual information of a quantum channel:

$$I_{\text{alt}}(\mathcal{N}) \equiv \max_{\rho^{XAB}} I(AX; B), \tag{12.127}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where the maximization is over states of the form ρ^{XAB}. Show that

$$I_{\text{alt}}(\mathcal{N}) = I(\mathcal{N}). \tag{12.128}$$

Exercise 12.4.2 Compute the mutual information of a dephasing channel with dephasing parameter p.

Exercise 12.4.3 Compute the mutual information of an erasure channel with erasure parameter ϵ.

Exercise 12.4.4 (Pure States are Sufficient) Show that it is sufficient to consider pure state $\phi^{AA'}$ for determining the mutual information of a quantum channel. That is, one does not need to consider mixed states $\rho^{AA'}$ in the optimization task. (Hint: use the spectral decomposition, the quantum data processing inequality, and apply the result of Exercise 12.4.1.)

12.4.2 Optimizing the Mutual Information of a Quantum Channel

We now show that the mutual information of a quantum channel is concave as a function of the input state. This result allows us to compute this quantity with standard convex optimization techniques.

Theorem 12.4.2. *The mutual information $I(A;B)$ is concave in the input state, in the sense that*

$$\sum_x p_X(x) I(A;B)_{\rho_x} \leq I(A;B)_\sigma, \tag{12.129}$$

where $\rho_x^{AB} \equiv \mathcal{N}^{A'\to B}(\phi_x^{AA'})$, $\sigma^{A'} \equiv \sum_x p_X(x)\rho_x^{A'}$, $\phi^{AA'}$ is a purification of $\sigma^{A'}$, and $\sigma^{AB} \equiv \mathcal{N}^{A'\to B}(\phi^{AA'})$.

Proof. Let ρ^{XABE} be the following classical-quantum state:

$$\rho^{XABE} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes U_{\mathcal{N}}^{A'\to BE}(\phi_x^{AA'}), \tag{12.130}$$

where $U_{\mathcal{N}}^{A'\to BE}$ is the isometric extension of the channel. Consider the following chain of inequalities:

$$\begin{aligned}
\sum_x p_X(x) I(A;B)_{\rho_x} &= I(A;B|X)_\rho && (12.131)\\
&= H(A|X)_\rho + H(B|X)_\rho - H(AB|X)_\rho && (12.132)\\
&= H(BE|X)_\rho + H(B|X)_\rho - H(E|X)_\rho && (12.133)\\
&= H(B|EX)_\rho + H(B|X)_\rho && (12.134)\\
&\leq H(B|E)_\rho + H(B)_\rho && (12.135)\\
&= H(B|E)_\sigma + H(B)_\sigma && (12.136)\\
&= I(A;B)_\sigma && (12.137)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first equality follows because the conditioning system X in $I(A;B|X)_\rho$ is classical. The second equality follows by expanding the quantum mutual information. The third equality follows because the state on ABE is pure when conditioned on X. The fourth equality follows from the definition of conditional quantum entropy. The inequality follows from strong subadditivity and concavity of quantum entropy. The equality follows by inspecting the definition of the state σ, and the final equality follows because the state is pure on systems ABE. □

12.5 Coherent Information of a Quantum Channel

This section presents an alternative, important measure of the ability of a quantum channel to preserve quantum correlations: the coherent information of the channel. The way we arrive at this measure is similar to how we did for the mutual information of a quantum channel. Alice prepares a pure state $\phi^{AA'}$ and inputs the A' system to a quantum channel $\mathcal{N}^{A'\to B}$. This transmission leads to a noisy state ρ^{AB} where

$$\rho^{AB} = \mathcal{N}^{A'\to B}\left(\phi^{AA'}\right). \tag{12.138}$$

The coherent information of the state that arises from the channel is as follows:

$$I(A\rangle B)_\rho \equiv H(B)_\rho - H(AB)_\rho, \tag{12.139}$$

leading to our next definition.

Definition 12.5.1 (Coherent Information of a Quantum Channel)**.** *The coherent information $Q(\mathcal{N})$ of a quantum channel is the maximum of the coherent information over all input states:*

$$Q(\mathcal{N}) \equiv \max_{\phi^{AA'}} I(A\rangle B)_\rho. \tag{12.140}$$

The coherent information of a quantum channel corresponds to an important operational task (perhaps the most important for quantum information). It is a good lower bound on the ultimate rate at which Alice can transmit quantum information to Bob, but it is actually equal to the quantum communication capacity of a quantum channel in some special cases. We prove these results rigorously in Chapter 23.

Exercise 12.5.1 Let $I_c(\rho,\mathcal{N})$ denote the coherent information of a channel $\mathcal{N}$ when state ρ is its input:

$$I_c(\rho,\mathcal{N}) \equiv H(\mathcal{N}(\rho)) - H(\mathcal{N}^c(\rho)), \tag{12.141}$$

where $\mathcal{N}^c$ is a channel complementary to the original channel $\mathcal{N}$. Show that

$$Q(\mathcal{N}) = \max_\rho I_c(\rho,\mathcal{N}). \tag{12.142}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

An equivalent way of writing the above expression on the RHS is

$$\max_{\phi^{AA'}} \Big[H(B)_{\psi} - H(E)_{\psi} \Big], \tag{12.143}$$

where $|\psi\rangle^{ABE} \equiv U_{\mathcal{N}}^{A'\to BE}|\phi\rangle^{AA'}$ and $U_{\mathcal{N}}^{A'\to BE}$ is the isometric extension of the channel $\mathcal{N}$.

The following property points out that the coherent information of a channel is always positive, even though the coherent information of any given state can sometimes be negative.

Property 12.5.1 (Non-negativity of Channel Coherent Information) The coherent information $Q(\mathcal{N})$ of a quantum channel $\mathcal{N}$ is non-negative:

$$Q(\mathcal{N}) \geq 0. \tag{12.144}$$

Proof. We can choose the input state $\phi^{AA'}$ to be a product state of the form $\psi^A \otimes \varphi^{A'}$. The coherent information of this state vanishes:

$$\begin{aligned} I(A\rangle B)_{\psi^A \otimes \mathcal{N}(\varphi^{A'})} &= H(B)_{\mathcal{N}(\varphi^{A'})} - H(AB)_{\psi^A \otimes \mathcal{N}(\varphi^{A'})} && (12.145)\\ &= H(B)_{\mathcal{N}(\varphi^{A'})} - H(A)_{\psi^A} - H(B)_{\mathcal{N}(\varphi^{A'})} && (12.146)\\ &= 0. && (12.147) \end{aligned}$$

The first equality follows by evaluating the coherent information for the product state. The second equality follows because the state on AB is product. The last equality follows because the state on A is pure. The above property then holds because the coherent information of a channel can only be greater than this amount, given that it involves a maximization over all input states and the above state is a particular input state. □

12.5.1 Additivity of Coherent Information for Degradable Channels

The coherent information of a quantum channel is generally not additive for arbitrary quantum channels. You might potentially view this situation as unfortunate, but it implies that quantum Shannon theory is a richer theory than its classical counterpart. Attempts to understand why and how this quantity is not additive have led to many breakthroughs (see Section 23.7).

Degradable quantum channels form a special class of channels for which the coherent information is additive. These channels have a property that is analogous to a property of the degraded wiretap channels from Section 12.2. To understand this property, recall that any quantum channel $\mathcal{N}^{A'\to B}$ has a complementary channel $(\mathcal{N}^c)^{A'\to E}$, realized by considering the isometric extension of the channel and tracing over Bob's system.

Definition 12.5.2 (Degradable Quantum Channel)**.** *A degradable quantum channel is one for which there exists a degrading map $\mathcal{T}^{B\to E}$ so that for any input state $\rho^{A'}$:*

$$(\mathcal{N}^c)^{A'\to E}\Big(\rho^{A'}\Big) = \mathcal{T}^{B\to E}\Big(\mathcal{N}^{A'\to B}\Big(\rho^{A'}\Big)\Big). \tag{12.148}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The intuition behind a degradable quantum channel is that the map from Alice to Eve is noisier than the map from Alice to Bob, in the sense that Bob can simulate the map to Eve by applying a noisy map to his system. We can show additivity of coherent information for these channels by exploiting a technique similar to that in the proof of additivity for degraded wiretap channels. The picture to consider for the analysis of additivity is the same as in Figure 12.5.

Theorem 12.5.1 (Additivity of the Coherent Information of Degradable Quantum Channels)**.** *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be any quantum channels that are degradable. Then the coherent information of the tensor product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is the sum of their individual coherent informations:*

$$Q(\mathcal{N}_1 \otimes \mathcal{N}_2) = Q(\mathcal{N}_1) + Q(\mathcal{N}_2). \tag{12.149}$$

Proof. We leave the proof of the inequality $Q(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq Q(\mathcal{N}_1) + Q(\mathcal{N}_2)$ as Exercise 12.5.3 below, and we prove the non-trivial inequality $Q(\mathcal{N}_1 \otimes \mathcal{N}_2) \leq Q(\mathcal{N}_1) + Q(\mathcal{N}_2)$ that holds when quantum channels $\mathcal{N}_1$ and $\mathcal{N}_2$ are degradable. Consider a pure state $\phi^{AA_1'A_2'}$ that serves as the input to the two quantum channels. Let $U_{\mathcal{N}_1}^{A_1' \to B_1 E_1}$ denote the isometric extension of the first channel and let $U_{\mathcal{N}_2}^{A_2' \to B_2 E_2}$ denote the isometric extension of the second channel. Let

$$\sigma^{AB_1E_1A_2'} \equiv U_{\mathcal{N}_1} \phi U_{\mathcal{N}_1}^\dagger, \tag{12.150}$$

$$\theta^{AA_1'B_2E_2} \equiv U_{\mathcal{N}_2} \phi U_{\mathcal{N}_2}^\dagger, \tag{12.151}$$

$$\rho^{AB_1E_1B_2E_2} \equiv (U_{\mathcal{N}_1} \otimes U_{\mathcal{N}_2}) \phi \Big(U_{\mathcal{N}_1}^\dagger \otimes U_{\mathcal{N}_2}^\dagger \Big). \tag{12.152}$$

We need to show that $Q(\mathcal{N}_1 \otimes \mathcal{N}_2) = Q(\mathcal{N}_1) + Q(\mathcal{N}_2)$ when both channels are degradable. Furthermore, let $\rho^{AB_1E_1B_2E_2}$ be the state that maximizes $Q(\mathcal{N}_1 \otimes \mathcal{N}_2)$. Consider the following chain of inequalities:

$$\begin{aligned}
Q(\mathcal{N}_1 \otimes \mathcal{N}_2) &= I(A\rangle B_1 B_2)_\rho && (12.153)\\
&= H(B_1B_2)_\rho - H(AB_1B_2)_\rho && (12.154)\\
&= H(B_1B_2)_\rho - H(E_1E_2)_\rho && (12.155)\\
&= H(B_1)_\rho - H(E_1)_\rho + H(B_2)_\rho - H(E_2)_\rho \\
&\quad - \Big[I(B_1;B_2)_\rho - I(E_1;E_2)_\rho \Big] && (12.156)\\
&\leq H(B_1)_\rho - H(E_1)_\rho + H(B_2)_\rho - H(E_2)_\rho && (12.157)\\
&= H(B_1)_\sigma - H(AA_2'B_1)_\sigma + H(B_2)_\theta - H(AA_1'B_2)_\theta && (12.158)\\
&= I(AA_2'\rangle B_1)_\sigma + I(AA_1'\rangle B_2)_\theta && (12.159)\\
&\leq Q(\mathcal{N}_1) + Q(\mathcal{N}_2). && (12.160)
\end{aligned}$$

The first equality follows from the definition of $Q(\mathcal{N}_1 \otimes \mathcal{N}_2)$ and because we set ρ to be the state that maximizes the tensor product channel coherent information. The second equality follows from the definition of coherent information, and the third equality follows because

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

the state ρ is pure on systems $AB_1E_1B_2E_2$. The fourth equality follows by expanding the entropies in the previous line. The first inequality (the crucial one) follows because there is a degrading channel from both B_1 to E_1 and B_2 to E_2, allowing us to apply the quantum data processing inequality twice to get $I(B_1;B_2)_\rho \geq I(E_1;E_2)_\rho$. The fifth equality follows because the entropies of ρ, σ, and θ on the given reduced systems are equal and because the state σ on systems $AA_2'B_1E_1$ is pure and the state θ on systems $AA_1'B_2E_2$ is pure. The last equality follows from the definition of coherent information, and the final inequality follows because the coherent informations are less than their respective maximizations over all possible states. □

Corollary 12.5.1. *The regularized coherent information of a degradable quantum channel is equal to its coherent information:*

$$Q_{reg}(\mathcal{N}) = Q(\mathcal{N}). \tag{12.161}$$

Proof. The proof of this property uses the same induction argument as in Corollary 12.1.1 and exploits the additivity property in Theorem 12.5.1 above. □

Exercise 12.5.2 Consider the quantum erasure channel where the erasure paramater ε is such that $0 \leq \varepsilon \leq 1/2$. Find the channel that degrades this one and compute the coherent information of the erasure channel as a function of ε.

Exercise 12.5.3 (Superadditivity of Coherent Information) Show that the coherent information of the tensor product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is never less than the sum of their individual coherent informations:

$$Q(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq Q(\mathcal{N}_1) + Q(\mathcal{N}_2). \tag{12.162}$$

Exercise 12.5.4 Prove using monotonicity of relative entropy that the coherent information is subadditive for a degradable channel:

$$Q(\mathcal{N}_1) + Q(\mathcal{N}_2) \leq Q(\mathcal{N}_1 \otimes \mathcal{N}_2) \tag{12.163}$$

Exercise 12.5.5 Consider a quantity known as the reverse coherent information:

$$Q_{\text{rev}}(\mathcal{N}) \equiv \max_{\phi^{AA'}} I(B\rangle A)_{\mathcal{N}(\phi^{AA'})}. \tag{12.164}$$

Show that the reverse coherent information is additive for any two quantum channels $\mathcal{N}_1$ and $\mathcal{N}_2$:

$$Q_{\text{rev}}(\mathcal{N}_1 \otimes \mathcal{N}_2) = Q_{\text{rev}}(\mathcal{N}_1) + Q_{\text{rev}}(\mathcal{N}_2). \tag{12.165}$$

12.5.2 Optimizing the Coherent Information of a Degradable Channel

We would like to determine how difficult it is to maximize the coherent information of a quantum channel. For general channels, this problem is difficult, but it turns out to

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

be straightforward for the class of degradable quantum channels. Theorem 12.5.2 below states an important property of the coherent information $Q(\mathcal{N})$ of a degradable quantum channel $\mathcal{N}$ that allows us to answer this question. The theorem states that the coherent information $Q(\mathcal{N})$ of a degradable quantum channel is a concave function of the input density operator $\rho^{A'}$ over which we maximize it. In particular, this result implies that the coherent information $Q(\mathcal{N})$ has a unique global maximum since the set of density operators is convex, and the optimization problem is therefore a straightforward computation that can exploit convex optimization methods. The below theorem exploits the characterization of the channel coherent information from Exercise 12.5.1.

Theorem 12.5.2. *Suppose that a quantum channel $\mathcal{N}$ is degradable. Then the coherent information $I_c(\rho, \mathcal{N})$ is concave in the input density operator:*

$$\sum_x p_X(x) I_c(\rho_x, \mathcal{N}) \leq I_c\left(\sum_x p_X(x)\rho_x, \mathcal{N}\right), \tag{12.166}$$

where $p_X(x)$ is a probability density function and each ρ_x is a density operator.

Proof. Consider the following states:

$$\sigma^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}(\rho_x), \tag{12.167}$$

$$\theta^{XE} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes (\mathcal{T} \circ \mathcal{N})(\rho_x), \tag{12.168}$$

where $\mathcal{T}$ is degrading map of the channel $\mathcal{N}$ so that

$$\mathcal{T} \circ \mathcal{N} = \mathcal{N}^c. \tag{12.169}$$

Then the following statements hold:

$$I(X;B)_\sigma \geq I(X;E)_\theta \tag{12.170}$$

$$\therefore \quad H(B)_\sigma - H(B|X)_\sigma \geq H(E)_\theta - H(E|X)_\theta \tag{12.171}$$

$$\therefore \quad H(B)_\sigma - H(E)_\theta \geq H(B|X)_\sigma - H(E|X)_\theta \tag{12.172}$$

$$\therefore \quad H\left(\mathcal{N}\left(\sum_x p_X(x)\rho_x\right)\right) - H\left(\mathcal{N}^c\left(\sum_x p_X(x)\rho_x\right)\right) \geq \sum_x p_X(x)(H(\mathcal{N}(\rho_x)) - H(\mathcal{N}^c(\rho_x))) \tag{12.173}$$

$$\therefore \quad I_c\left(\sum_x p_X(x)\rho_x, \mathcal{N}\right) \geq \sum_x p_X(x) I_c(\rho_x, \mathcal{N}) \tag{12.174}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first statement is the crucial one and follows from the quantum data processing inequality and the fact that the map $\mathcal{T}$ degrades Bob's state to Eve's state. The second and third statements follow from the definition of quantum mutual information and rearranging entropies. The fourth statement follows by plugging in the density operators into the entropies in the previous statement. The final statement follows from the alternate definition of coherent information in Exercise 12.5.1. □

12.6 Private Information of a Quantum Channel

The private information of a quantum channel is the last information measure that we consider. Alice would like to establish classical correlations with Bob, but does not want the environment of the channel to have access to these classical correlations. The ensemble that she prepares is similar to the one we considered for the Holevo information. The expected density operator of the ensemble she prepares is a classical-quantum state of the form:

$$\rho^{XA'} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^{A'}. \tag{12.175}$$

Sending the A' system through the isometric extension $U_{\mathcal{N}}^{A'\to BE}$ of a quantum channel $\mathcal{N}$ leads to a state ρ^{XBE}. A good measure of the private classical correlations that she can establish with Bob is the difference of the classical correlations she can establish with Bob, less the classical correlations that Eve can obtain:

$$I(X;B)_\rho - I(X;E)_\rho, \tag{12.176}$$

leading to our next definition (Chapter 22 discusses the operational task corresponding to this information quantity).

Definition 12.6.1 (Private Information of a Quantum Channel)**.** *The private information $P(\mathcal{N})$ of a quantum channel $\mathcal{N}$ is the maximization over all of Alice's input preparations:*

$$P(\mathcal{N}) \equiv \max_{\rho^{XA'}} I(X;B)_\rho - I(X;E)_\rho. \tag{12.177}$$

Property 12.6.1 The private information $P(\mathcal{N})$ of a quantum channel $\mathcal{N}$ is non-negative:

$$P(\mathcal{N}) \geq 0. \tag{12.178}$$

Proof. We can choose the input state $\rho^{XA'}$ to be a state of the form $|0\rangle\langle 0|^X \otimes \psi^{A'}$, where $\psi^{A'}$ is pure. The private information of this state vanishes:

$$I(X;B)_{|0\rangle\langle 0|\otimes\mathcal{N}(\psi)} - I(X;E)_{|0\rangle\langle 0|\otimes\mathcal{N}^c(\psi)} = 0. \tag{12.179}$$

The equality follows just by evaluating both mutual informations for the above state. The above property then holds because the private information of a channel can only be greater than this amount, given that it involves a maximization over all input states and the above state is a particular input state. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The regularized private information is as follows:

$$P_{\text{reg}}(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n} P\big(\mathcal{N}^{\otimes n}\big). \tag{12.180}$$

12.6.1 The Relationship of Private Information with Coherent Information

The private information of a quantum channel bears a special relationship to that channel's coherent information. It is always at least as great as the coherent information of the channel and is equal to it for certain channels. The following theorem states the former inequality, and the next theorem states the equivalence for degradable quantum channels.

Theorem 12.6.1. *The private information $P(\mathcal{N})$ of any quantum channel $\mathcal{N}$ is at least as large as its coherent information $Q(\mathcal{N})$:*

$$Q(\mathcal{N}) \leq P(\mathcal{N}). \tag{12.181}$$

Proof. We can see this relation through a few steps. Consider a pure state $\phi^{AA'}$ that maximizes the coherent information $Q(\mathcal{N})$, and let ϕ^{ABE} denote the state that arises from sending the A' system through the isometric extension $U_{\mathcal{N}}^{A'\to BE}$ of the channel $\mathcal{N}$. Let $\phi^{A'}$ denote the reduction of this state to the A' system. Suppose that it admits the following spectral decomposition:

$$\phi^{A'} = \sum_x p_X(x) |\phi_x\rangle\langle\phi_x|^{A'}. \tag{12.182}$$

We can create an augmented classical-quantum state that correlates a classical variable with the index x:

$$\sigma^{XA'} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes |\phi_x\rangle\langle\phi_x|^{A'}. \tag{12.183}$$

Let σ^{XBE} denote the state that results from from sending the A' system through the isometric extension $U_{\mathcal{N}}^{A'\to BE}$ of the channel $\mathcal{N}$. Then the following chain of inequalities holds:

$$\begin{aligned}
Q(\mathcal{N}) &= I(A\rangle B)_\phi && (12.184)\\
&= H(B)_\phi - H(E)_\phi && (12.185)\\
&= H(B)_\sigma - H(E)_\sigma && (12.186)\\
&= H(B)_\sigma - H(B|X)_\sigma - H(E)_\sigma + H(B|X)_\sigma && (12.187)\\
&= I(X;B)_\sigma - H(E)_\sigma + H(E|X)_\sigma && (12.188)\\
&= I(X;B)_\sigma - I(X;E)_\sigma && (12.189)\\
&\leq P(\mathcal{N}). && (12.190)
\end{aligned}$$

The first equality follows from evaluating the coherent information of the state ϕ^{ABE} that maximizes the coherent information of the channel. The second equality follows because the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

state ϕ^{ABE} is pure. The third equality follows from the definition of σ^{XBE} in (12.183) and its relation to ϕ^{ABE}. The fourth equality follows by adding and subtracting $H(B|X)_\sigma$, and the next one follows from the definition of the mutual information $I(X;B)_\sigma$ and the fact that the state of σ^{XBE} on systems B and E is pure when conditioned on X. The last equality follows from the definition of the mutual information $I(X;E)_\sigma$. The final inequality follows because the state σ^{XBE} is a particular state of the form in (12.175), and $P(\mathcal{N})$ involves a maximization over all states of that form. □

Theorem 12.6.2. *Suppose that a quantum channel $\mathcal{N}$ is degradable. Then its private information $P(\mathcal{N})$ is equal to its coherent information $Q(\mathcal{N})$:*

$$P(\mathcal{N}) = Q(\mathcal{N}). \tag{12.191}$$

Proof. We prove the inequality $P(\mathcal{N}) \leq Q(\mathcal{N})$ for degradable quantum channels because we have already proven that $Q(\mathcal{N}) \leq P(\mathcal{N})$ for any quantum channel $\mathcal{N}$. Consider a classical-quantum state ρ^{XBE} that arises from transmitting the A' system of the state in (12.175) through the isometric extension $U_{\mathcal{N}}^{A'\to BE}$ of the channel. Suppose further that this state maximizes $P(\mathcal{N})$. We can take the spectral decomposition of each $\rho_x^{A'}$ in the ensemble to be as follows:

$$\rho_x^{A'} = \sum_y p_{Y|X}(y|x)\psi_{x,y}^{A'}, \tag{12.192}$$

where each state $\psi_{x,y}^{A'}$ is pure. We can construct the following extension of the state ρ^{XBE} as follows:

$$\sigma^{XYBE} \equiv \sum_{x,y} p_{Y|X}(y|x)p_X(x)|x\rangle\langle x|^X \otimes |y\rangle\langle y|^Y \otimes U_{\mathcal{N}}^{A'\to BE}(\psi_{x,y}^{A'}). \tag{12.193}$$

Then the following chain of inequalities holds:

$$\begin{aligned}
P(\mathcal{N}) &= I(X;B)_\rho - I(X;E)_\rho && (12.194)\\
&= I(X;B)_\sigma - I(X;E)_\sigma && (12.195)\\
&= I(XY;B)_\sigma - I(Y;B|X)_\sigma - [I(XY;E)_\sigma - I(Y;E|X)_\sigma] && (12.196)\\
&= I(XY;B)_\sigma - I(XY;E)_\sigma - [I(Y;B|X)_\sigma - I(Y;E|X)_\sigma] && (12.197)
\end{aligned}$$

The first equality follows because from the definition of $P(\mathcal{N})$ and because we set ρ to be the state that maximizes it. The second equality follows because $\rho^{XBE} = \mathrm{Tr}_Y\{\sigma^{XYBE}\}$. The third equality follows from the chain rule for quantum mutual information. The fourth equality follows from a rearrangement of entropies. Continuing,

$$\begin{aligned}
&\leq I(XY;B)_\sigma - I(XY;E)_\sigma && (12.198)\\
&= H(B)_\sigma - H(B|XY)_\sigma - H(E)_\sigma + H(E|XY)_\sigma && (12.199)\\
&= H(B)_\sigma - H(B|XY)_\sigma - H(E)_\sigma + H(B|XY)_\sigma && (12.200)\\
&= H(B)_\sigma - H(E)_\sigma && (12.201)\\
&\leq Q(\mathcal{N}). && (12.202)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

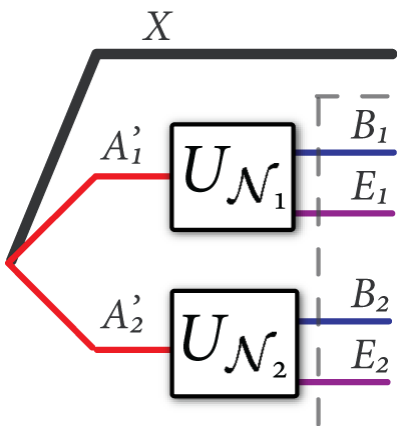

Figure 12.6: The above figure displays the scenario for determining whether the private information of two quantum channels $\mathcal{N}_1$ and $\mathcal{N}_2$ is additive. The question of additivity is equivalent to the possibility of quantum correlations between channel inputs being able to enhance the private information of two quantum channels. The result proved in Theorem 12.6.3 is that the private information is additive for any two degradable quantum channels, so that quantum correlations cannot enhance it in this case.

The first inequality (the crucial one) follows because there is a degrading channel from B to E and because the conditioning system X is classical, allowing us to apply the quantum data processing inequality $I(Y;B|X)_\sigma \geq I(Y;E|X)_\sigma$. The second equality is a rewriting of entropies, the third follows because the state of σ on systems B and E is pure when conditioned on classical systems X and Y, and the fourth follows by canceling entropies. The last inequality follows because the entropy difference $H(B)_\sigma - H(E)_\sigma$ is less than the maximum of that difference over all possible states. □

12.6.2 Additivity of the Private Information of Degradable Channels

The private information of general quantum channels is not additive, but it is so in the case of degradable quantum channels. The method of proof is somewhat similar to that in the proof of Theorem 12.5.1, essentially exploiting the degradability property. Figure 12.6 illustrates the setting to consider for additivity of the private information.

Theorem 12.6.3 (Additivity of Private Information of Degradable Quantum Channels)**.** *Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be any quantum channels that are degradable. Then the private information of the tensor product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is the sum of their individual private informations:*

$$P(\mathcal{N}_1 \otimes \mathcal{N}_2) = P(\mathcal{N}_1) + P(\mathcal{N}_2). \tag{12.203}$$

Furthermore, it holds that

$$P(\mathcal{N}_1 \otimes \mathcal{N}_2) = Q(\mathcal{N}_1 \otimes \mathcal{N}_2) = Q(\mathcal{N}_1) + Q(\mathcal{N}_2). \tag{12.204}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Proof. We first prove the more trivial inequality $P(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq P(\mathcal{N}_1) + P(\mathcal{N}_2)$. Let $\rho^{X_1 A_1'}$ and $\sigma^{X_2 A_2'}$ be states of the form in (12.175) that maximize the respective private informations $P(\mathcal{N}_1)$ and $P(\mathcal{N}_2)$. Let $\theta^{X_1 X_2 A_1' A_2'}$ be the tensor product of these two states: $\theta = \rho \otimes \sigma$. Let $\rho^{X_1 B_1 E_1}$ and $\sigma^{X_2 B_2 E_2}$ be the states that arise from sending $\rho^{X_1 A_1'}$ and $\sigma^{X_2 A_2'}$ through the respective isometric extensions $U_{\mathcal{N}}^{A_1' \to B_1 E_1}$ and $U_{\mathcal{N}}^{A_2' \to B_2 E_2}$. Let $\theta^{X_1 X_2 B_1 B_2 E_1 E_2}$ be the state that arises from sending $\theta^{X_1 X_2 A_1' A_2'}$ through the tensor product channel $U_{\mathcal{N}}^{A_1' \to B_1 E_1} \otimes U_{\mathcal{N}}^{A_2' \to B_2 E_2}$. Then

$$
\begin{aligned}
& P(\mathcal{N}_1) + P(\mathcal{N}_2) \\
& = I(X_1; B_1)_\rho - I(X_1; E_1)_\rho + I(X_2; B_2)_\sigma - I(X_2; E_2)_\sigma && (12.205) \\
& = I(X_1; B_1)_\theta - I(X_1; E_1)_\theta + I(X_2; B_2)_\theta - I(X_2; E_2)_\theta && (12.206) \\
& = H(B_1)_\theta - H(B_1|X_1)_\theta - H(E_1)_\theta + H(E_1|X_1)_\theta \\
& \quad + H(B_2)_\theta - H(B_2|X_2)_\theta - H(E_2)_\theta + H(E_2|X_2)_\theta && (12.207) \\
& = H(B_1 B_2)_\theta - H(B_1 B_2|X_1 X_2)_\theta \\
& \quad - H(E_1 E_2)_\theta - H(E_1 E_2|X_1 X_2)_\theta && (12.208) \\
& = I(X_1 X_2; B_1 B_2)_\theta - I(X_1 X_2; E_1 E_2)_\theta && (12.209) \\
& \leq P(\mathcal{N}_1 \otimes \mathcal{N}_2). && (12.210)
\end{aligned}
$$

The first equality follows from the definition of the private informations $P(\mathcal{N}_1)$ and $P(\mathcal{N}_2)$ and by evaluating them on the respective states $\rho^{X_1 A_1'}$ and $\sigma^{X_2 A_2'}$ that maximize them. The second equality follows because the reduced state of $\theta^{X_1 X_2 B_1 B_2 E_1 E_2}$ on systems X_1, B_1, and E_1 is equal to $\rho^{X_1 B_1 E_1}$, and the reduced state of $\theta^{X_1 X_2 B_1 B_2 E_1 E_2}$ on systems X_2, B_2, and E_2 is equal to $\sigma^{X_2 B_2 E_2}$. The third equality follows by expanding the mutual informations. The fourth equality follows because the state $\theta^{X_1 X_2 B_1 B_2 E_1 E_2}$ is product with respect to the systems $X_1 B_1 E_1$ and $X_2 B_2 E_2$. The fifth equality follows from straightforward entropic manipulations, and the final inequality follows because the state $\theta^{X_1 X_2 B_1 B_2 E_1 E_2}$ is a particular state of the form needed in the maximization of the private information of the tensor product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$. We now prove the inequality $P(\mathcal{N}_1 \otimes \mathcal{N}_2) \leq P(\mathcal{N}_1) + P(\mathcal{N}_2)$. Let $\rho^{X A_1' A_2'}$ be the state that maximizes $P(\mathcal{N}_1 \otimes \mathcal{N}_2)$ where

$$
\rho^{X A_1' A_2'} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes \rho_x^{A_1' A_2'}, \tag{12.211}
$$

and let $\rho^{X B_1 B_2 E_1 E_2}$ be the state that arises from sending $\rho^{X A_1' A_2'}$ through the tensor product channel $U_{\mathcal{N}}^{A_1' \to B_1 E_1} \otimes U_{\mathcal{N}}^{A_2' \to B_2 E_2}$. Consider a spectral decomposition of each state $\rho_x^{A_1' A_2'}$:

$$
\rho_x^{A_1' A_2'} = \sum_y p_{Y|X}(y|x) \psi_{x,y}^{A_1' A_2'}, \tag{12.212}
$$

where each state $\psi_{x,y}^{A_1' A_2'}$ is pure. Let $\sigma^{XY A_1' A_2'}$ be an extension of $\rho^{X A_1' A_2'}$ where

$$
\rho^{XY A_1' A_2'} \equiv \sum_{x,y} p_{Y|X}(y|x) p_X(x) |x\rangle\langle x|^X \otimes |y\rangle\langle y|^Y \otimes \psi_{x,y}^{A_1' A_2'}, \tag{12.213}
$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and let $\sigma^{XYB_1E_1B_2E_2}$ be the state that arises from sending $\sigma^{XYA_1'A_2'}$ through the tensor product channel $U_{\mathcal{N}}^{A_1'\to B_1E_1} \otimes U_{\mathcal{N}}^{A_2'\to B_2E_2}$. Consider the following chain of inequalities:

$$\begin{aligned}
&P(\mathcal{N}_1 \otimes \mathcal{N}_2) && (12.214)\\
&= I(X; B_1B_2)_\rho - I(X; E_1E_2)_\rho && (12.215)\\
&= I(X; B_1B_2)_\sigma - I(X; E_1E_2)_\sigma && (12.216)\\
&= I(XY; B_1B_2)_\sigma - I(XY; E_1E_2)_\sigma \\
&\quad - [I(Y; B_1B_2|X)_\sigma - I(Y; E_1E_2|X)_\sigma] && (12.217)\\
&\leq I(XY; B_1B_2)_\sigma - I(XY; E_1E_2)_\sigma && (12.218)\\
&= H(B_1B_2)_\sigma - H(B_1B_2|XY)_\sigma - H(E_1E_2)_\sigma + H(E_1E_2|XY)_\sigma && (12.219)\\
&= H(B_1B_2)_\sigma - H(B_1B_2|XY)_\sigma - H(E_1E_2)_\sigma + H(B_1B_2|XY)_\sigma && (12.220)\\
&= H(B_1B_2)_\sigma - H(E_1E_2)_\sigma && (12.221)\\
&= H(B_1)_\sigma - H(E_1)_\sigma + H(B_2)_\sigma - H(E_2)_\sigma \\
&\quad - [I(B_1; B_2)_\sigma - I(E_1; E_2)_\sigma] && (12.222)\\
&\leq H(B_1)_\sigma - H(E_1)_\sigma + H(B_2)_\sigma - H(E_2)_\sigma && (12.223)\\
&\leq Q(\mathcal{N}_1) + Q(\mathcal{N}_2) && (12.224)\\
&= P(\mathcal{N}_1) + P(\mathcal{N}_2). && (12.225)
\end{aligned}$$

The first equality follows from the definition of $P(\mathcal{N}_1 \otimes \mathcal{N}_2)$ and evaluating it on the state ρ that maximizes it. The second equality follows because the state $\sigma^{XYB_1E_1B_2E_2}$ is equal to the state $\rho^{XB_1E_1B_2E_2}$ after tracing out the system Y. The third equality follows from the chain rule for mutual information: $I(XY; B_1B_2) = I(Y; B_1B_2|X) + I(X; B_1B_2)$. It holds that $I(Y; B_1B_2|X)_\sigma \geq I(Y; E_1E_2|X)_\sigma$ because the conditioning system X is classical and there is a degrading channel from B_1 to E_1 and from B_2 to E_2. Then the first inequality follows because $I(Y; B_1B_2|X)_\sigma - I(Y; E_1E_2|X)_\sigma \geq 0$. The fourth equality follows by expanding the mutual informations, and the fifth equality follows because the state σ on systems $B_1B_2E_1E_2$ is pure when conditioning on the classical systems X and Y. The sixth equality follows from algebra, and the seventh follows by rewriting the entropies. It holds that $I(B_1; B_2)_\sigma \geq I(E_1; E_2)_\sigma$ because there is a degrading channel from B_1 to E_1 and from B_2 to E_2. Then the inequality follows because $I(B_1; B_2)_\sigma - I(E_1; E_2)_\sigma \geq 0$. The third inequality follows because the entropy difference $H(B_i) - H(E_i)$ is always less than the coherent information of the channel, and the final equality follows because the coherent information of a channel is equal to its private information when the channel is degradable (Theorem 12.6.2). □

Corollary 12.6.1. *Suppose that a quantum channel $\mathcal{N}$ is degradable. Then the regularized private information $P_{reg}(\mathcal{N})$ of the channel is equal to its private information $P(\mathcal{N})$:*

$$P_{reg}(\mathcal{N}) = P(\mathcal{N}). \tag{12.226}$$

Proof. The proof follows by the same induction argument as in Corollary 12.1.1 and by exploiting the result of Theorem 12.6.3 and the fact that the tensor power channel $\mathcal{N}^{\otimes n}$ is degradable if the original channel $\mathcal{N}$ is. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

12.7 Summary

We conclude this chapter with a table that summarizes the main results regarding the mutual information of a classical channel $I(p_{Y|X})$, the private information of a classical wiretap channel $P(p_{Y,Z|X})$, the Holevo information of a quantum channel $\chi(\mathcal{N})$, the mutual information of a quantum channel $I(\mathcal{N})$, the coherent information of a quantum channel $Q(\mathcal{N})$, and the private information of a quantum channel $P(\mathcal{N})$. The table exploits the following definitions:

$$\rho^{XA'} \equiv \sum p_X(x)|x\rangle\langle x|^X \otimes \phi_x^{A'}, \tag{12.227}$$

$$\sigma^{XA'} \equiv \sum p_X(x)|x\rangle\langle x|^X \otimes \rho_x^{A'}. \tag{12.228}$$

Quantity	Input	Output	Formula	Single-letter
$I(p_{Y\|X})$	$p_X(x)$	$p_X(x)p_{Y\|X}(y\|x)$	$\max_{p_X(x)} I(X;Y)$	all channels
$P(p_{Y,Z\|X})$	$p_X(x)$	$p_X(x)p_{Y,Z\|X}(y,z\|x)$	$\max_{p_X(x)} I(X;Y) - I(X;Z)$	degradable
$\chi(\mathcal{N})$	$\rho^{XA'}$	$\mathcal{N}^{A'\to B}(\rho^{XA'})$	$\max_\rho I(X;B)$	many channels
$I(\mathcal{N})$	$\phi^{AA'}$	$\mathcal{N}^{A'\to B}(\phi^{AA'})$	$\max_\phi I(A;B)$	all channels
$Q(\mathcal{N})$	$\phi^{AA'}$	$\mathcal{N}^{A'\to B}(\phi^{AA'})$	$\max_\phi I(A\rangle B)$	degradable
$P(\mathcal{N})$	$\sigma^{XA'}$	$U_{\mathcal{N}}^{A'\to BE}(\sigma^{XA'})$	$\max_\sigma I(X;B) - I(X;E)$	degradable

12.8 History and Further Reading

The book of Boyd and Vandenberghe is useful for the theory and practice of convex optimization [45], which is helpful for computing capacity formulas. Holevo [144], Schumacher, and Westmoreland [219] provided an operational interpretation of the Holevo information of a quantum channel. Shor showed the additivity of the Holevo information for entanglement-breaking channels [226]. Adami and Cerf introduced the mutual information of a quantum channel, and they proved several of its important properties that appear in this chapter: non-negativity, additivity, and concavity [5]. Bennett *et al.* later gave an operational interpretation for this information quantity as the entanglement-assisted classical capacity of a quantum channel [33, 34]. Lloyd [185], Shor [227], and Devetak [68] gave increasingly rigorous proofs that the coherent information of a quantum channel is an achievable rate for quantum communication. Devetak and Shor showed that the coherent information of a quantum channel is additive for degradable channels [73]. Yard *et al.* proved that the coherent information of a quantum channel is a concave function of the input state whenever the channel is degradable [269]. García-Patrón *et al.* and Devetak *et al.* both discussed the reverse coherent information of a quantum channel and showed that it is additive for all quantum channels [100, 72]. Devetak [68] and Cai *et al.* [51] independently introduced the private classical capacity of a quantum channel, and both papers proved that it is an achievable rate for private classical communication over a quantum channel. Smith showed

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

that the private classical information is additive and equal to the coherent information for degradable quantum channels [230].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 13

Classical Typicality

This chapter begins our first technical foray into the asymptotic theory of information. We start with the classical setting in an effort to build up our intuition of asymptotic behavior before delving into the asymptotic theory of quantum information.

The central concept of this chapter is the asymptotic equipartition property. The name of this property may sound somewhat technical at first, but it is merely an application of the law of large numbers to a sequence drawn independently and identically from a distribution $p_X(x)$ for some random variable X. The asymptotic equipartition property reveals that we can divide sequences into two classes when their length becomes large: those that are overwhelmingly likely to occur and those that are overwhelmingly likely not to occur. The sequences that are likely to occur are the *typical* sequences, and the ones that are not likely to occur are the *atypical* sequences. Additionally, the size of the set of typical sequences is exponentially smaller than the size of the set of all sequences whenever the random variable generating the sequences is not uniform. These properties are an example of a more general mathematical phenomenon known as "measure concentration," in which a smooth function over a high-dimensional space or over a large number of random variables tends to concentrate around a constant value with high probability.

The asymptotic equipartition property immediately leads to the intuition behind Shannon's scheme for compressing classical information. The scheme first generates a realization of a random sequence and asks the question: Is the produced sequence typical or atypical? If it is typical, compress it. Otherwise, throw it away. The error probability of this compression scheme is non-zero for any fixed length of a sequence, but it vanishes in the asymptotic limit because the probability of the sequence being in the typical set converges to one, while the probability that it is in the atypical set converges to zero. This compression scheme has a straightforward generalization to the quantum setting, where we wish to compress qubits instead of classical bits.

The bulk of this chapter is here to present the many technical details needed to make rigorous statements in the asymptotic theory of information. We begin with an example, follow with the formal definition of a typical sequence and a typical set, and prove the three important properties of a typical set. We then discuss other forms of typicality such

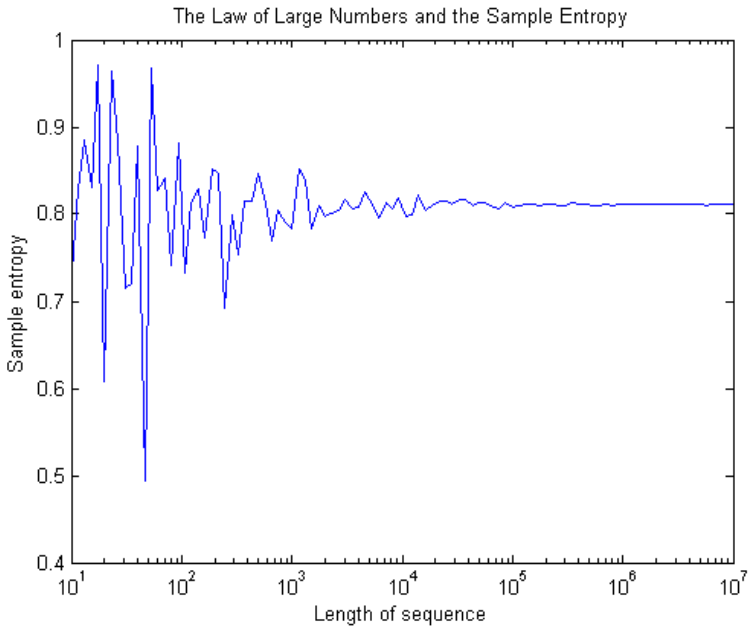

Figure 13.1: The above figure depicts the sample entropy of a realization of a random binary sequence as a function of its length. The source is a binary random variable with distribution $(\frac{3}{4}, \frac{1}{4})$. For the realizations generated, the sample entropy of the sequences is converging to the true entropy of the source.

as joint typicality and conditional typicality. These other notions turn out to be useful for proving Shannon's classical capacity theorem as well (recall that Shannon's theorem gives the ultimate rate at which a sender can transmit classical information over a classical channel to a receiver). We also introduce the method of types, which is a powerful technique in classical information theory, and apply this method in order to develop a stronger notion of typicality. The chapter then features a development of the strong notions of joint and conditional typicality and ends with a concise proof of Shannon's important channel capacity theorem.

13.1 An Example of Typicality

Suppose that Alice possesses a binary random variable X that takes the value zero with probability $\frac{3}{4}$ and the value one with probability $\frac{1}{4}$. Such a random source might produce the following sequence:

$$0110001101, \tag{13.1}$$

if we generate ten realizations of it. The probability that such a sequence occurs is

$$\left(\frac{1}{4}\right)^5 \left(\frac{3}{4}\right)^5, \tag{13.2}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

determined simply by counting the number of ones and zeros in the above sequence and by applying the independent and identically distributed (IID) property of the source.

The *information content* of the above sequence is the negative logarithm of its probability divided by its length:

$$-\frac{5}{10}\log\left(\frac{3}{4}\right)-\frac{5}{10}\log\left(\frac{1}{4}\right)\approx 1.207. \tag{13.3}$$

We also refer to this quantity as *the sample entropy*. The true entropy of the source is

$$-\frac{1}{4}\log\left(\frac{1}{4}\right)-\frac{3}{4}\log\left(\frac{3}{4}\right)\approx 0.8113. \tag{13.4}$$

We would expect that the sample entropy of a random sequence tends to approach the true entropy as its size increases because the number of zeros should be approximately $n(3/4)$ and the number of ones should be approximately $n(1/4)$ according to the law of large numbers.

Another sequence of length 100 might be as follows:

$$\begin{aligned}&0000000010001000100000000000011001101000000000100000\\&0000011010100100000001000000100000001000001000100000,\end{aligned} \tag{13.5}$$

featuring 81 zeros and 19 ones. Its sample entropy is

$$-\frac{81}{100}\log\left(\frac{3}{4}\right)-\frac{19}{100}\log\left(\frac{1}{4}\right)\approx 0.7162. \tag{13.6}$$

The above sample entropy is closer to the true entropy in (13.4) than the sample entropy of the previous sequence, but it still deviates significantly from it.

Figure 13.1 continues this game by generating random sequences according to the distribution $\left(\frac{3}{4},\frac{1}{4}\right)$, and the result is that a concentration around the true entropy begins to occur around $n\approx 10^6$. That is, it becomes highly likely that the sample entropy of a random sequence is close to the true entropy if we increase the length of the sequence, and this holds for the realizations generated in Figure 13.1.

13.2 Weak Typicality

This first section generalizes the example from the introduction to an arbitrary discrete, finite-valued random variable. Our first notion of typicality is the same discussed in the example—we define a sequence to be typical if its sample entropy is close to the true entropy of the random variable that generates it. This notion of typicality is known as *weak typicality*. Section 13.7 introduces another notion of typicality that implies weak typicality, but the implication does not hold in the other direction. For this reason, we distinguish the two different notions of typicality as weak typicality and strong typicality.

Suppose that a random variable X takes values in an alphabet $\mathcal{X}$ with cardinality $|\mathcal{X}|$. Let us label the symbols in the alphabet as a_1, a_2, $\ldots$, $a_{|\mathcal{X}|}$. An independent and identically

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

distributed (IID) information source samples *independently* from the distribution of random variable X and emits n realizations $x_1, \ldots, x_n$. Let $X^n \equiv X_1 \cdots X_n$ denote the n random variables that describe the information source, and let $x^n \equiv x_1 \cdots x_n$ denote an emitted realization of X^n. The probability $p_{X^n}(x^n)$ of a particular string x^n is as follows:

$$p_{X^n}(x^n) \equiv p_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \tag{13.7}$$

and $p_{X^n}(x^n)$ factors as follows because the source is IID:

$$p_{X^n}(x^n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n) = p_X(x_1) \cdots p_X(x_n) = \prod_{i=1}^{n} p_X(x_i). \tag{13.8}$$

Roughly speaking, we expect a long string x^n to contain about $np_X(a_1)$ occurrences of symbol a_1, $np_X(a_2)$ occurrences of symbol a_2, etc., when n is large. The probability that the source emits a particular string x^n is approximately

$$p_{X^n}(x^n) = p_X(x_1) \cdots p_X(x_n) \approx p_X(a_1)^{np_X(a_1)} \cdots p_X\left(a_{|\mathcal{X}|}\right)^{np_X\left(a_{|\mathcal{X}|}\right)}, \tag{13.9}$$

and the information content of a given string is thus roughly

$$-\frac{1}{n} \log(p_{X^n}(x^n)) \approx -\sum_{i=1}^{|\mathcal{X}|} p_X(a_i) \log(p_X(a_i)) = H(X). \tag{13.10}$$

The above intuitive argument shows that the information content divided by the length of the sequence is roughly equal to the entropy in the limit of large n. It then makes sense to think of this quantity as the *sample entropy* of the sequence x^n.

Definition 13.2.1 (Sample Entropy)**.** *The sample entropy $\overline{H}(x^n)$ of a sequence x^n is as follows:*

$$\overline{H}(x^n) \equiv -\frac{1}{n} \log(p_{X^n}(x^n)). \tag{13.11}$$

This definition of sample entropy leads us to our first important definitions in asymptotic information theory.

Definition 13.2.2 (Typical Sequence)**.** *A sequence x^n is δ-typical if its sample entropy $\overline{H}(x^n)$ is δ-close to the entropy $H(X)$ of random variable X, where this random variable is the source of the sequence.*

Definition 13.2.3 (Typical Set)**.** *The δ-typical set $T_\delta^{X^n}$ is the set of all δ-typical sequences x^n:*

$$T_\delta^{X^n} \equiv \left\{x^n : \left|\overline{H}(x^n) - H(X)\right| \leq \delta\right\}. \tag{13.12}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

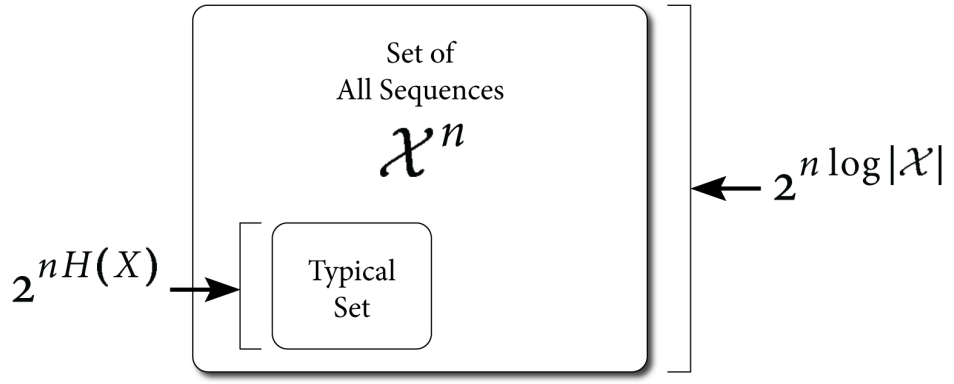

Figure 13.2: The above figure depicts the idea that the typical set is exponentially smaller than the set of all sequences because $|\mathcal{X}|^n = 2^{n \log |\mathcal{X}|} > 2^{nH(X)}$ whenever X is not a uniform random variable. Yet, this exponentially small set contains nearly all of the probability.

13.3 Properties of the Typical Set

The set of typical sequences enjoys three useful and beautifully surprising properties that occur when we step into the "land of large numbers." We can summarize these properties as follows: the typical set contains almost all the probability, yet it is exponentially smaller than the set of all sequences, and each typical sequence has almost uniform probability. Figure 13.2 attempts to depict the main idea of the typical set.

Property 13.3.1 (Unit Probability) The typical set asymptotically has probability one. So as n becomes large, it is highly likely that a source emits a typical sequence. We formally state this property as follows:

$$\forall \epsilon > 0 \qquad \Pr\left\{X^n \in T_\delta^{X^n}\right\} = \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n) \geq 1 - \epsilon \quad \text{for sufficiently large } n. \tag{13.13}$$

Property 13.3.2 (Exponentially Small Cardinality) The number $\left|T_\delta^{X^n}\right|$ of δ-typical sequences is exponentially smaller than the total number $|\mathcal{X}|^n$ of sequences for every random variable X besides the uniform random variable. We formally state this property as follows:

$$\left|T_\delta^{X^n}\right| \leq 2^{n(H(X)+\delta)}. \tag{13.14}$$

We can also lower bound the size of the δ-typical set when n is sufficiently large:

$$\forall \epsilon > 0 \qquad \left|T_\delta^{X^n}\right| \geq (1-\epsilon) 2^{n(H(X)-\delta)} \quad \text{for sufficiently large } n. \tag{13.15}$$

Property 13.3.3 (Equipartition) The probability of a particular δ-typical sequence x^n is approximately uniform:

$$2^{-n(H(X)+\delta)} \leq p_{X^n}(x^n) \leq 2^{-n(H(X)-\delta)}. \tag{13.16}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

This last property is the "equipartition" in "asymptotic equipartition property" because all typical sequences occur with nearly the same probability when n is large.

The size $\left|T_\delta^{X^n}\right|$ of the δ-typical set is approximately equal to the total number $|\mathcal{X}|^n$ of sequences only when random variable X is uniform because $H(X) = \log|\mathcal{X}|$ and thus

$$\left|T_\delta^{X^n}\right| \leq 2^{n(H(X)+\delta)} = 2^{n(\log|\mathcal{X}|+\delta)} = |\mathcal{X}|^n \cdot 2^{n\delta} \simeq |\mathcal{X}|^n. \tag{13.17}$$

13.3.1 Proofs of Typical Set Properties

Proof of the Unit Probability Property (Property 13.3.1). The weak law of large numbers states that the sample mean converges in probability to the expectation. More precisely, consider a sequence of IID random variables $X_1, \ldots, X_n$ that each have expectation μ. The sample average of this sequence is as follows:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{13.18}$$

The formal statement of the law of large numbers is

$$\forall \epsilon, \delta > 0 \;\; \exists n_0 : \forall n > n_0 \qquad \Pr\left\{\left|\overline{X} - \mu\right| < \delta\right\} > 1 - \epsilon. \tag{13.19}$$

We can now consider the sequence of random variables $-\log(p_X(X_1)), \ldots, -\log(p_X(X_n))$. The sample average of this sequence is equal to the sample entropy of X^n:

$$-\frac{1}{n}\sum_{i=1}^{n} \log(p_X(X_i)) = -\frac{1}{n}\log(p_{X^n}(X^n)) \tag{13.20}$$

$$= \overline{H}(X^n). \tag{13.21}$$

Recall from (10.3) that the expectation of the random variable $-\log(p_X(X))$ is equal to the Shannon entropy:

$$\mathbb{E}_X\{-\log(p_X(X))\} = H(X). \tag{13.22}$$

Then we can apply the law of large numbers and find that

$$\forall \epsilon, \delta > 0 \;\; \exists n_0 : \forall n > n_0 \qquad \Pr\left\{\left|\overline{H}(X^n) - H(X)\right| < \delta\right\} > 1 - \epsilon. \tag{13.23}$$

The event $\left\{\left|\overline{H}(X^n) - H(X)\right| < \delta\right\}$ is precisely the condition for a random sequence X^n to be in the typical set $T_\delta^{X^n}$, and the probability of this event goes to one as n becomes large. □

Proof of the Exponentially Small Cardinality Property (Property 13.3.2). Consider the following chain of inequalities:

$$1 = \sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n) \geq \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n)$$

$$\geq \sum_{x^n \in T_\delta^{X^n}} 2^{-n(H(X)+\delta)} = 2^{-n(H(X)+\delta)}\left|T_\delta^{X^n}\right|. \tag{13.24}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first inequality uses the fact that the probability of the typical set is smaller than the probability of the set of all sequences. The second inequality uses the equipartition property of typical sets (proved below). After rearranging the leftmost side of (13.24) with its rightmost side, we find that

$$\left|T_\delta^{X^n}\right| \leq 2^{n(H(X)+\delta)}. \tag{13.25}$$

The second part of the property follows because the "unit probability" property holds for sufficiently large n. Then the following chain of inequalities holds:

$$\begin{aligned} 1-\epsilon \leq \Pr\left\{X^n \in T_\delta^{X^n}\right\} &= \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n) \\ &\leq \sum_{x^n \in T_\delta^{X^n}} 2^{-n(H(X)-\delta)} = 2^{-n(H(X)-\delta)}\left|T_\delta^{X^n}\right|. \end{aligned} \tag{13.26}$$

We can then bound the size of the typical set as follows:

$$\left|T_\delta^{X^n}\right| \geq 2^{n(H(X)-\delta)}(1-\epsilon), \tag{13.27}$$

for any $\epsilon > 0$ and sufficiently large n. □

Proof of the Equipartition Property (Property 13.3.3). The property follows immediately by manipulating the definition of a typical set. □

13.4 Application of Typical Sequences: Shannon Compression

The above three properties of typical sequences immediately give our first application in asymptotic information theory. It is Shannon's compression protocol, which is a scheme for compressing the output of an information source.

We begin by defining the information processing task and a corresponding (n, R, ϵ) source code. It is helpful to recall the picture in Figure 2.1. An information source outputs a sequence x^n drawn independently according to the distribution of some random variable X. A sender Alice encodes this sequence according to some encoding map E where

$$E : \mathcal{X}^n \to \{0,1\}^{nR}. \tag{13.28}$$

The encoding takes elements from the set $\mathcal{X}^n$ of all sequences to a set $\{0,1\}^{nR}$ of size 2^{nR}. She then transmits the codewords over nR uses of a noiseless classical channel. Bob decodes according to some decoding map $D : \{0,1\}^{nR} \to \mathcal{X}^n$. The probability of error for an (n, R, ϵ) source code is

$$p(e) \equiv \Pr\{(D \circ E)(X^n) \neq X^n\} \leq \epsilon. \tag{13.29}$$

The rate of the source code is the number of channel uses divided by the length of the sequence, and it is equal to R for the above scheme. A particular compression rate R is *achievable* if there exists an $(n, R+\delta, \epsilon)$ source code for all $\epsilon, \delta > 0$ and all sufficiently large n. We can now state Shannon's lossless compression theorem.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

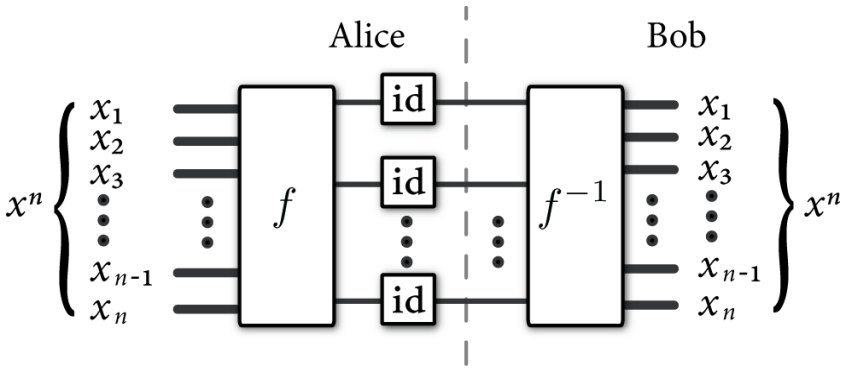

Figure 13.3: Shannon's scheme for the compression of classical data. The encoder f is a map from the typical set to a set of binary sequences of size $\approx 2^{nH(X)}$ where $H(X)$ is the entropy of the information source. The map f is invertible on the typical set but maps an atypical sequence to a constant. Alice then transmits the compressed data over $\approx nH(X)$ uses of a noiseless classical channel. The inverse map f^{-1} (the decoder) is the inverse of f on the typical set and decodes to some error sequence otherwise.

Theorem 13.4.1 (Shannon Compression)**.** *The entropy of the source is the smallest achievable rate for compression:*

$$\inf\{R : R \text{ is achievable}\} = H(X). \tag{13.30}$$

The proof of this theorem consists of two parts, traditionally called the direct coding theorem and the converse theorem. The direct coding theorem is the direction LHS $\leq$ RHS—the proof exhibits a coding scheme with an achievable rate and demonstrates that its rate converges to the entropy in the asymptotic limit. The converse theorem is the direction LHS $\geq$ RHS and is a statement of optimality—it proves that any coding scheme with rate below the entropy is not achievable. The proofs of each part are usually completely different. We employ typical sequences and their properties for proving a direct coding theorem, while the converse part resorts to information inequalities from Chapter 10.[1] For now, we prove the direct coding theorem and hold off on the converse part until we reach Schumacher compression for quantum information in Chapter 17. Our main goal here is to illustrate a simple application of typical sequences, and we can wait on the converse part because Shannon compression is a special case of Schumacher compression.

The idea behind the proof of the direct coding theorem is simple: just compress the typical sequences and throw away the rest. A code of this form succeeds with asymptotically vanishing probability of error because the typical set asymptotically has all of the probability. Since we are only concerned with error probabilities in communication protocols, it makes sense that we should only be keeping track of a set where all of the probability concentrates. We can formally state the proof as follows. Pick an $\epsilon > 0$, a $\delta > 0$, and a sufficiently large n

[1]The direct part of a quantum coding theorem can employ the properties of typical subspaces (discussed in Chapter 14), and the proof of a converse theorem for quantum information usually employs the quantum information inequalities from Chapter 11.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

such that Property 13.3.1 holds. Consider that Property 13.3.2 then holds so that the size of the typical set is no larger than $2^{n[H(X)+\delta]}$. We choose the encoding to be a function f that maps a typical sequence to a binary sequence in $\{0,1\}^{nR}\backslash e_0$, where $R = H(X) + \delta$ and e_0 is some error symbol in $\{0,1\}^{nR}$. We define f to map any atypical sequence to e_0. This scheme gives up on encoding the atypical sequences because they have vanishingly small probability. We define the decoding operation to be the inverse of f on the typical set, while mapping to some fixed sequence $x^n \in \mathcal{X}^n$ if the received symbol is e_0. This scheme has probability of error less than ϵ, by considering Property 13.3.1. Figure 13.3 depicts this coding scheme.

Shannon's scheme for compression suffers from a problem that plagues all results in classical and quantum information theory. The proof guarantees that there exists a scheme that can compress at the rate of entropy in the asymptotic limit. But the complexity of encoding and decoding is far from practical—without any further specification of the encoding, it could require resources that are prohibitively exponential in the size of the sequence.

The above scheme certainly gives an achievable rate for compression of classical information, but how can we know that it is optimal? The converse theorem addresses this point (recall that a converse theorem gives a sense of optimality for a particular protocol) and completes the operational interpretation of the entropy as the fundamental limit on the compressibility of classical information. For now, we do not prove a converse theorem and instead choose to wait until we cover Schumacher compression because its converse proof applies to Shannon compression as well.

13.5 Weak Joint Typicality

Joint typicality is a concept similar to typicality, but the difference is that it applies to any two random variables X and Y. That is, there are analogous notions of typicality for the joint random variable (X, Y).

Definition 13.5.1 (Joint Sample Entropy)**.** *Consider n independent realizations $x^n = x_1 \cdots x_n$ and $y^n = y_1 \cdots y_n$ of respective random variables X and Y. The sample joint entropy $\overline{H}(x^n, y^n)$ of these two sequences is*

$$\overline{H}(x^n, y^n) \equiv -\frac{1}{n} \log(p_{X^n,Y^n}(x^n, y^n)), \tag{13.31}$$

where we assume that the joint distribution $p_{X^n,Y^n}(x^n, y^n)$ has the IID property:

$$p_{X^n,Y^n}(x^n, y^n) \equiv p_{X,Y}(x_1, y_1) \cdots p_{X,Y}(x_n, y_n). \tag{13.32}$$

This notion of joint sample entropy immediately leads to the following definition of joint typicality. Figure 13.4 attempts to depict the notion of joint typicality.

Definition 13.5.2 (Jointly Typical Sequence)**.** *Two sequences x^n, y^n are δ-jointly-typical if their sample joint entropy $\overline{H}(x^n, y^n)$ is δ-close to the joint entropy $H(X, Y)$ of random variables X and Y and if both x^n and y^n are marginally typical.*

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

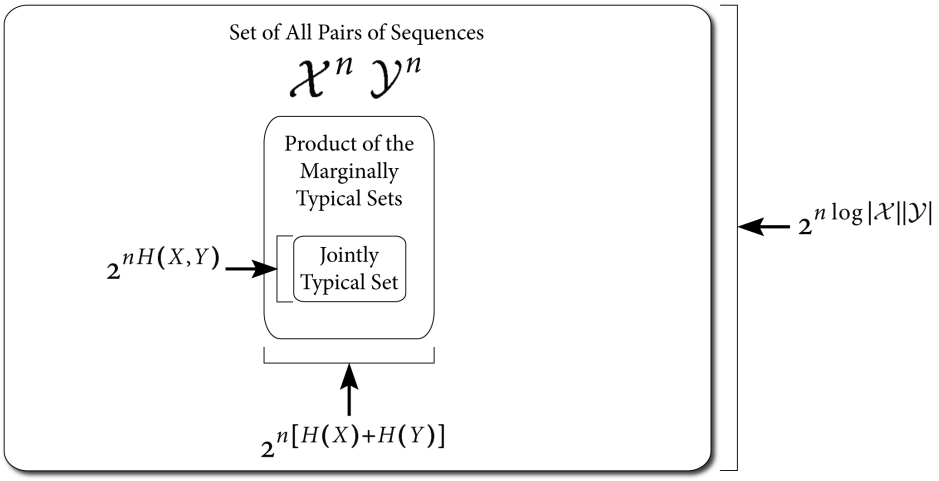

Figure 13.4: A depiction of the jointly typical set. Some sequence pairs (x^n, y^n) are such that x^n is typical or such that y^n is typical, but fewer are such that the pair is jointly typical. The jointly typical set has size roughly equal to $2^{nH(X,Y)}$, which is smaller than the Cartesian product of the marginally typical sets if random variables X and Y are not independent.

Definition 13.5.3 (Jointly Typical Set)**.** *The δ-jointly typical set $T_\delta^{X^nY^n}$ consists of all δ-jointly typical sequences:*

$$T_\delta^{X^nY^n} \equiv \left\{x^n, y^n : \left|\overline{H}(x^n, y^n) - H(X,Y)\right| \leq \delta, \quad x^n \in T_\delta^{X^n}, \quad y^n \in T_\delta^{Y^n}\right\}. \tag{13.33}$$

The extra conditions on the marginal sample entropies are necessary to have a sensible definition of joint typicality. That is, it does not necessarily follow that the marginal sample entropies are close to the marginal true entropies if the joint ones are close, but it intuitively makes sense that this condition should hold. Thus, we add these extra conditions to the definition of jointly typical sequences. Later, we find in Section 13.7 that the intuitive implication holds (it is not necessary to include the marginals) when we employ a stronger definition of typicality.

13.5.1 Properties of the Jointly Typical Set

The set $T_\delta^{X^nY^n}$ of jointly typical sequences enjoys three properties similar to what we have seen in Section 13.2, and the proofs of these properties are identical to those in Section 13.2.

Property 13.5.1 (Unit Probability) The jointly typical set asymptotically has probability one. So as n becomes large, it is highly likely that a source emits a jointly typical sequence. We formally state this property as follows:

$$\forall \epsilon > 0 \qquad \Pr\left\{X^nY^n \in T_\delta^{X^nY^n}\right\} \geq 1 - \epsilon \quad \text{for sufficiently large } n. \tag{13.34}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Property 13.5.2 (Exponentially Small Cardinality) The number $\left|T_\delta^{X^nY^n}\right|$ of δ-jointly typical sequences is exponentially smaller than the total number $(|\mathcal{X}||\mathcal{Y}|)^n$ of sequences for any joint random variable (X, Y) that is not uniform. We formally state this property as follows:

$$\left|T_\delta^{X^nY^n}\right| \leq 2^{n(H(X,Y)+\delta)}. \tag{13.35}$$

We can also lower bound the size of the δ-jointly typical set when n is sufficiently large:

$$\forall \epsilon > 0 \qquad \left|T_\delta^{X^nY^n}\right| \geq (1-\epsilon)2^{n(H(X,Y)-\delta)} \quad \text{for sufficiently large } n. \tag{13.36}$$

Property 13.5.3 (Equipartition) The probability of a particular δ-jointly typical sequence x^ny^n is approximately uniform:

$$2^{-n(H(X,Y)+\delta)} \leq p_{X^n,Y^n}(x^n, y^n) \leq 2^{-n(H(X,Y)-\delta)}. \tag{13.37}$$

Exercise 13.5.1 Prove the above three properties of the jointly typical set.

The above three properties may be similar to what we have seen before, but there is another interesting property of jointly typical sequences that we give below. It states that two sequences drawn independently according to the marginal distributions $p_X(x)$ and $p_Y(y)$ are jointly typical according to the joint distribution $p_{X,Y}(x, y)$ with probability $\approx 2^{-nI(X;Y)}$. This property gives a simple interpretation of the mutual information that is related to its most important operational interpretation as the classical channel capacity discussed briefly in Section 2.2.

Property 13.5.4 (Probability of Joint Typicality) Consider two independent random variables $\tilde{X}^n$ and $\tilde{Y}^n$ whose respective probability density functions $p_{\tilde{X}^n}(x^n)$ and $p_{\tilde{Y}^n}(y^n)$ are equal to the marginal densities of the joint density $p_{X^n,Y^n}(x^n, y^n)$:

$$\left(\tilde{X}^n, \tilde{Y}^n\right) \sim p_{X^n}(x^n)p_{Y^n}(y^n). \tag{13.38}$$

Then we can bound the probability that two random sequences $\tilde{X}^n$ and $\tilde{Y}^n$ are in the jointly typical set $T_\delta^{X^nY^n}$:

$$\Pr\left\{\left(\tilde{X}^n, \tilde{Y}^n\right) \in T_\delta^{X^nY^n}\right\} \leq 2^{-n(I(X;Y)-3\delta)}. \tag{13.39}$$

Exercise 13.5.2 Prove Property 13.5.4. (Hint: Consider that

$$\Pr\left\{\left(\tilde{X}^n, \tilde{Y}^n\right) \in T_\delta^{X^nY^n}\right\} = \sum_{x^n,y^n \in T_\delta^{X^nY^n}} p_{X^n}(x^n)p_{Y^n}(y^n), \tag{13.40}$$

and use the properties of typical and jointly typical sets to bound this probability.)

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

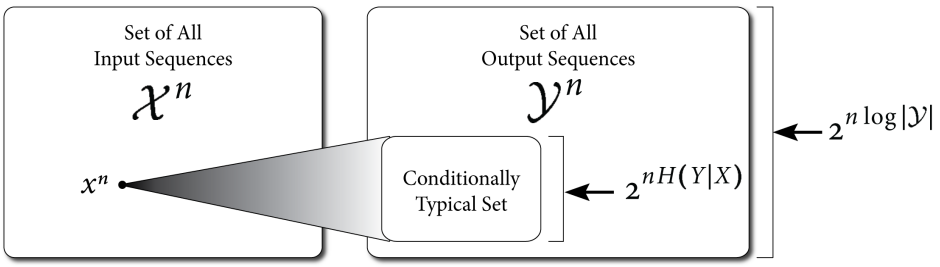

Figure 13.5: The notion of the conditionally typical set. A typical sequence x^n in $T_\delta^{X^n}$ maps stochastically through many instantiations of a conditional distribution $p_{Y|X}(y|x)$ to some sequence y^n. It is overwhelmingly likely that y^n is in a conditionally typical set $T_\delta^{Y^n|x^n}$ when n becomes large. This conditionally typical set has size around $2^{nH(Y|X)}$. It contains nearly all of the probability but is exponentially smaller than the set of all sequences $\mathcal{Y}^n$.

13.6 Weak Conditional Typicality

Conditional typicality is a property that we expect to hold for any two random sequences—it is also a useful tool in the proofs of coding theorems. Suppose two random variables X and Y have respective alphabets $\mathcal{X}$ and $\mathcal{Y}$ and a joint distribution $p_{X,Y}(x,y)$. We can factor the joint distribution $p_{X,Y}(x,y)$ as the product of a marginal distribution $p_X(x)$ and a conditional distribution $p_{Y|X}(y|x)$, and this factoring leads to a particular way that we can think about generating realizations of the joint random variable. We can consider random variable Y to be a noisy version of X, where we first generate a realization x of the random variable X according to the distribution $p_X(x)$ and follow by generating a realization y of the random variable Y according to the conditional distribution $p_{Y|X}(y|x)$.

Suppose that we generate n independent realizations of random variable X to obtain the sequence $x^n = x_1 \cdots x_n$. We then record these values and use the conditional distribution $p_{Y|X}(y|x)$ n times to generate n independent realizations of random variable Y. Let $y^n = y_1 \cdots y_n$ denote the resulting sequence.

Definition 13.6.1 (Conditional Sample Entropy)**.** *The conditional sample entropy* $\overline{H}(y^n|x^n)$ *of two sequences* x^n *and* y^n *is*

$$\overline{H}(y^n|x^n) = -\frac{1}{n} \log p_{Y^n|X^n}(y^n|x^n), \tag{13.41}$$

where

$$p_{Y^n|X^n}(y^n|x^n) \equiv p_{Y|X}(y_1|x_1) \cdots p_{Y|X}(y_n|x_n). \tag{13.42}$$

Definition 13.6.2 (Conditionally Typical Set)**.** *Suppose that a sequence* x^n *is a typical sequence in* $T_\delta^{X^n}$ *and* y^n *is a typical sequence in* $T_\delta^{Y^n}$*. The* δ*-conditionally typical set* $T_\delta^{Y^n|x^n}$ *consists of all sequences whose conditional sample entropy is* δ*-close to the true conditional entropy:*

$$T_\delta^{Y^n|x^n} \equiv \left\{ y^n : \left| \overline{H}(y^n|x^n) - H(Y|X) \right| \leq \delta \right\}. \tag{13.43}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

A different way to define the conditionally typical set is as follows.

Remark 13.6.1 The δ-conditionally typical set $T_\delta^{Y^n|x^n}$ corresponding to the sequence x^n consists of all sequences y^n that are jointly typical with x^n:

$$T_\delta^{Y^n|x^n} \equiv \left\{ y^n : (x^n, y^n) \in T_\delta^{X^n,Y^n} \right\}. \tag{13.44}$$

The appearance of marginal typicality in the above definitions may again seem somewhat strange, but they are there because it makes sense that the sequence y^n should be typical if it is conditionally typical (we thus impose this constraint in the definition). This property does not necessarily follow from the weak notion of conditional typicality, but it does follow without any imposed constraints from a stronger notion of conditional typicality that we give in Section 13.9.

13.6.1 Properties of the Conditionally Typical Set

The set $T_\delta^{Y^n|x^n}$ of conditionally typical sequences enjoys properties similar to what we have seen before, and we list them for completeness.

Property 13.6.1 (Unit Probability) The set $T_\delta^{Y^n|x^n}$ asymptotically has probability one when the sequence x^n is random. So as n becomes large, it is highly likely that random sequences Y^n and X^n are such that Y^n is a conditionally typical sequence. We formally state this property as follows:

$$\forall \epsilon > 0 \qquad \mathbb{E}_{X^n} \left\{ \Pr_{Y^n|X^n} \left\{ Y^n \in T_\delta^{Y^n|X^n} \right\} \right\} \geq 1 - \epsilon \quad \text{for sufficiently large } n. \tag{13.45}$$

Property 13.6.2 (Exponentially Small Cardinality) The number $\left| T_\delta^{Y^n|x^n} \right|$ of δ-conditionally typical sequences is exponentially smaller than the total number $|\mathcal{Y}|^n$ of sequences for any conditional random variable $Y|X$ that is not uniform. We formally state this property as follows:

$$\left| T_\delta^{Y^n|x^n} \right| \leq 2^{n(H(Y|X)+\delta)}. \tag{13.46}$$

We can also lower bound the expected size of the δ-conditionally typical set when n is sufficiently large and x^n is a random sequence:

$$\forall \epsilon > 0 \qquad \mathbb{E}_{X^n} \left\{ \left| T_\delta^{Y^n|X^n} \right| \right\} \geq (1 - \epsilon) 2^{n(H(Y|X)-\delta)} \quad \text{for sufficiently large } n. \tag{13.47}$$

Property 13.6.3 (Equipartition) The probability of a given δ-conditionally typical sequence y^n (corresponding to the sequence x^n) is approximately uniform:

$$2^{-n(H(Y|X)+\delta)} \leq p_{Y^n|X^n}(y^n|x^n) \leq 2^{-n(H(Y|X)-\delta)}. \tag{13.48}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

In summary, averaged over realizations of the random variable X^n, the conditionally typical set $T_\delta^{Y^n|X^n}$ has almost all the probability, and its size is exponentially smaller than the size of the set of all sequences. For each realization of X^n, each δ-conditionally typical sequence has an approximate uniform probability of occurring.

Our last note on the weak conditionally typical set is that there is a subtlety in the statement of Property 13.6.1 that allows for a relatively straightforward proof. This subtlety is that we average over the sequence X^n well, and this allows one to exploit the extra randomness to simplify the proof. We do not impose such a constraint later on in Section 13.9 where we introduce the notion of a strong conditionally typical sequence. We instead impose the constraint that the sequence x^n is a strongly typical sequence, and this property is sufficient to prove that similar properties hold for a strong conditionally typical set.

We now prove the first property. Consider that

$$\begin{aligned}
&\mathbb{E}_{X^n}\left\{ \Pr_{Y^n|X^n}\left\{ Y^n \in T_\delta^{Y^n|X^n} \right\} \right\} \\
&= \sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n) \sum_{y^n \in T_\delta^{Y^n|x^n}} p_{Y^n|X^n}(y^n|x^n) && (13.49) \\
&\geq \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n) \sum_{y^n \in T_\delta^{Y^n|x^n}} p_{Y^n|X^n}(y^n|x^n) && (13.50) \\
&= \sum_{x^n \in T_\delta^{X^n}} \sum_{y^n \in T_\delta^{Y^n|x^n}} p_{X^n,Y^n}(x^n,y^n) && (13.51) \\
&\geq 1-\epsilon. && (13.52)
\end{aligned}$$

The first equality follows by definition. The first inequality follows because the probability mass of the set $\mathcal{X}^n$ can only be larger than the probability mass in the typical set $T_\delta^{X^n}$. The last inequality follows because the conditions $\left|\overline{H}(x^n) - H(X)\right| \leq \delta$ and $\left|\overline{H}(y^n|x^n) - H(Y|X)\right| \leq \delta$ imply

$$\left|\overline{H}(x^n,y^n) - H(X,Y)\right| \leq \delta' \tag{13.53}$$

for some δ', for which we then have the law of large numbers to obtain this final bound.

Exercise 13.6.1 Prove that the last two properties hold for the weak conditionally typical set.

13.7 Strong Typicality

In the development in the previous sections, we showed how the law of large numbers is the underpinning method to prove many of the interesting results regarding typical sequences. These results are satisfactory and provide an intuitive notion of typicality through the idea of the sample entropy approaching the true entropy for sufficiently long sequences.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

It is possible to develop a stronger notion of typicality with a different definition. Instead of requiring that the sample entropy of a random sequence is close to the true entropy of a distribution for sufficiently long sequences, strong typicality requires that the empirical distribution or relative frequency of symbols of a random sequence has a small deviation from the true probability distribution for sufficiently large sequence length.

We begin with a simple example to help illustrate this stronger notion of typicality. Suppose that we generate a binary sequence IID according to the distribution $p(0) = 1/4$ and $p(1) = 3/4$. Such a random generation process could lead to the following sequence:

$$0110111010. \tag{13.54}$$

Rather than computing the sample entropy of this sequence and comparing it with the true entropy, we can count the number of zeros or ones that appear in the sequence and compare their normalizations with the true distribution of the information source. For the above example, the number of zeros is equal to 4, and the number of ones (the Hamming weight of the sequence) is equal to 6:

$$N(0 \mid 0110111010) = 4, \qquad N(1 \mid 0110111010) = 6. \tag{13.55}$$

We can compute the empirical distribution of this sequence by normalizing the above numbers by the length of the sequence:

$$\frac{1}{10}N(0 \mid 0110111010) = \frac{2}{5}, \qquad \frac{1}{10}N(1 \mid 0110111010) = \frac{3}{5}. \tag{13.56}$$

This empirical distribution deviates from the true distribution by the following amount

$$\max\left\{\left|\frac{1}{4} - \frac{2}{5}\right|, \left|\frac{3}{4} - \frac{3}{5}\right|\right\} = \frac{3}{20}, \tag{13.57}$$

which is a fairly significant deviation. Though, suppose that the length of the sequence grows large enough that the law of large numbers comes into play. We would then expect it to be highly likely that the empirical distribution of a random sequence does not deviate much from the true distribution, and the law of large numbers again gives a theoretical underpinning for this intuition. This example gives the essence of strong typicality.

We wish to highlight another important aspect of the above example. The particular sequence in (13.54) has a Hamming weight of six, but this sequence is not the only one with this Hamming weight. By a simple counting argument, there are $\binom{10}{6} - 1 = 209$ other sequences with the same length and Hamming weight. That is, all these other sequences have the same empirical distribution and thus have the same deviation from the true distribution as the original sequence in (13.54). We say that all these sequences are in the same "type class," which simply means that they have the same empirical distribution. The type class is thus an equivalence class on sequences where the equivalence relation is the empirical distribution of the sequence.

We mention a few interesting properties of the type class before giving more formal definitions. We can partition the set of all possible sequences according to type classes.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Consider that the set of all binary sequences of length ten has size 2^{10}. There is one sequence with all zeros, $\binom{10}{1}$ sequences with Hamming weight one, $\binom{10}{2}$ sequences with Hamming weight two, etc. The binomial theorem guarantees that the total number of sequences is equal to the number of sequences in all of the type classes:

$$2^{10} = \sum_{i=0}^{10} \binom{10}{i}. \tag{13.58}$$

Suppose now that we generate ten IID realizations of the Bernoulli distribution $p(0) = 1/4$ and $p(1) = 3/4$. Without knowing anything else, our best description of the distribution of the random sequence is

$$p(x_1, \ldots, x_{10}) = p(x_1) \cdots p(x_{10}), \tag{13.59}$$

where $x_1, \ldots, x_{10}$ are different realizations of the binary random variable. But suppose that a third party tells us the Hamming weight w_0 of the generated sequence. This information allows us to update our knowledge of the distribution of the sequence, and we can say that any sequence with Hamming weight not equal to w_0 has zero probability. All the sequences with the same Hamming weight have the same distribution because we generated the sequence in an IID way, and each sequence with Hamming weight w_0 has a uniform distribution after renormalizing. Thus, conditioned on the Hamming weight w_0, our best description of the distribution of the random sequence is

$$p(x_1, \ldots, x_{10}|w_0) = \begin{cases} 0 & : \quad w(x_1, \ldots, x_{10}) \neq w_0 \\ \binom{10}{w_0}^{-1} & : \quad w(x_1, \ldots, x_{10}) = w_0 \end{cases}, \tag{13.60}$$

where w is a function that gives the Hamming weight of a binary sequence. This property has important consequences for asymptotic information processing because it gives us a way to extract uniform randomness from an IID distribution, and we later see that it has applications in several quantum information processing protocols as well.

13.7.1 Types and Strong Typicality

We now formally develop the notion of a type and strong typicality. Let x^n denote a sequence $x_1 x_2 \ldots x_n$, where each x_i belongs to the alphabet $\mathcal{X}$. Let $|\mathcal{X}|$ be the cardinality of $\mathcal{X}$. Let $N(x|x^n)$ be the number of occurrences of the symbol $x \in \mathcal{X}$ in the sequence x^n.

Definition 13.7.1 (Type)**.** *The type or empirical distribution t_{x^n} of a sequence x^n is a probability mass function whose elements are $t_{x^n}(x)$ where*

$$t_{x^n}(x) \equiv \frac{1}{n} N(x|x^n). \tag{13.61}$$

Definition 13.7.2 (Strongly Typical Set)**.** *The δ-strongly typical set $T_\delta^{X^n}$ is the set of all sequences with an empirical distribution $\frac{1}{n}N(x|x^n)$ that has maximum deviation δ from the*

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

true distribution $p_X(x)$. Furthermore, the empirical distribution $\frac{1}{n}N(x|x^n)$ of any sequence in $T_\delta^{X^n}$ vanishes for any letter x for which $p_X(x) = 0$:

$$T_\delta^{X^n} \equiv \left\{ x^n : \forall x \in \mathcal{X},\ \left| \frac{1}{n}N(x|x^n) - p_X(x) \right| \leq \delta \text{ if } p_X(x) > 0,\ \text{else } \frac{1}{n}N(x|x^n) = 0 \right\}. \tag{13.62}$$

The extra condition where $\frac{1}{n}N(x|x^n) = 0$ when $p_X(x) = 0$ is a somewhat technical condition, nevertheless intuitive, that is necessary to prove the three desired properties for the strongly typical set. Also, we are using the same notation $T_\delta^{X^n}$ to indicate both the weakly and strongly typical set, but which one is appropriate should be clear from context, or we will explicitly indicate which one we are using.

The notion of type class becomes useful for us in our later developments—it is simply a way for grouping together all the sequences with the same empirical distribution. Its most important use is as a way for obtaining a uniform distribution from an arbitrary IID distribution (recall that we can do this by conditioning on a particular type).

Definition 13.7.3 (Type Class)**.** *Let $T_t^{X^n}$ denote the type class of a particular type t. The type class $T_t^{X^n}$ is the set of all sequences with length n and type t:*

$$T_t^{X^n} \equiv \{x^n \in \mathcal{X}^n : t^{x^n} = t\}. \tag{13.63}$$

Property 13.7.1 (Bound on the Number of Types) The number of types for a given sequence length n containing symbols from an alphabet $\mathcal{X}$ is exactly equal to

$$\binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1}. \tag{13.64}$$

A good upper bound on the number of types is

$$(n+1)^{|\mathcal{X}|}. \tag{13.65}$$

Proof. The number of types is equivalent to the number of ways that the symbols in a sequence of length n can form $|\mathcal{X}|$ distinct groups. Consider the following visual aid:

$$\bullet\bullet\bullet\bullet\bullet \,|\, \bullet\bullet\bullet\bullet\bullet\bullet |\, \bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet \,|\, \bullet\bullet\bullet\bullet. \tag{13.66}$$

We can think of the number of types as the number of different ways of arranging $|\mathcal{X}| - 1$ vertical bars to group the n dots into $|\mathcal{X}|$ distinct groups. The upper bound follows from a simple argument. The number of types is the number of different ways that $|\mathcal{X}|$ positive numbers can sum to n. Overestimating the count, we can choose the first number in $n+1$ different ways (it can be any number from 0 to n), and we can choose the $|\mathcal{X}| - 1$ other numbers in $n+1$ different ways. Multiplying all of these possibilities together gives an upper bound $(n+1)^{|\mathcal{X}|}$ on the number of types. This bound illustrates that the number of types is only *polynomial* in the length n of the sequence (compare with the total number $|\mathcal{X}|^n$ of sequences of length n being exponential in the length of the sequence). □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Definition 13.7.4 (Typical Type)**.** *Let $p_X(x)$ denote the true probability distribution of symbols x in the alphabet $\mathcal{X}$. For $\delta > 0$, let τ_δ denote the set of all typical types that have maximum deviation δ from the true distribution $p_X(x)$:*

$$\tau_\delta \equiv \{t : \forall x \in \mathcal{X},\ |t(x) - p_X(x)| \leq \delta \text{ if } p_X(x) > 0 \text{ else } t(x) = 0\}. \tag{13.67}$$

We can then equivalently define the set of strongly δ-typical sequences of length n as a union over all the type classes of the typical types in τ_δ:

$$T_\delta^{X^n} = \bigcup_{t \in \tau_\delta} T_t^{X^n}. \tag{13.68}$$

13.7.2 Properties of the Strongly Typical Set

The strongly typical set enjoys many useful properties (similar to the weakly typical set).

Property 13.7.2 (Unit Probability) The strongly typical set asymptotically has probability one. So as n becomes large, it is highly likely that a source emits a strongly typical sequence. We formally state this property as follows:

$$\forall \epsilon > 0 \qquad \Pr\left\{X^n \in T_\delta^{X^n}\right\} \geq 1 - \epsilon \quad \text{for sufficiently large } n. \tag{13.69}$$

Property 13.7.3 (Exponentially Small Cardinality) The number $\left|T_\delta^{X^n}\right|$ of δ-typical sequences is exponentially smaller than the total number $|\mathcal{X}|^n$ of sequences for most random variables X. We formally state this property as follows:

$$\left|T_\delta^{X^n}\right| \leq 2^{n(H(X)+c\delta)}, \tag{13.70}$$

where c is some positive constant. We can also lower bound the size of the δ-typical set when n is sufficiently large:

$$\forall \epsilon > 0 \quad \left|T_\delta^{X^n}\right| \geq (1-\epsilon)2^{n(H(X)-c\delta)} \quad \text{for sufficiently large } n \text{ and some constant } c. \tag{13.71}$$

Property 13.7.4 (Equipartition) The probability of a given δ-typical sequence x^n occurring is approximately uniform:

$$2^{-n(H(X)+c\delta)} \leq p_{X^n}(x^n) \leq 2^{-n(H(X)-c\delta)}. \tag{13.72}$$

This last property of strong typicality demonstrates that it implies weak typicality up to an irrelevant constant c.

13.7.3 Proofs of the Properties of the Strongly Typical Set

Proof of the Unit Probability Property (Property 13.7.2). The proof proceeds similarly to the proof of the unit probability property for the weakly typical set. The law of large numbers states that the sample mean of a random sequence converges in probability to the expectation

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

of the random variable from which we generate the sequence. So consider a sequence of IID random variables $X_1, \ldots, X_n$ where each random variable in the sequence has expectation μ. The sample average of this sequence is as follows:

$$\overline{X} \equiv \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{13.73}$$

The precise statement of the weak law of large numbers is that

$$\forall \epsilon, \delta > 0 \;\; \exists n_0 : \forall n > n_0 \qquad \Pr\left\{ \left| \overline{X} - \mu \right| > \delta \right\} < \epsilon. \tag{13.74}$$

We can now consider the indicator random variables $I(X_1 = a), \ldots, I(X_n = a)$. The sample mean of a random sequence of indicator variables is equal to the empirical distribution $N(a|X^n)/n$:

$$\frac{1}{n} \sum_{i=1}^{n} I(X_i = a) = \frac{1}{n} N(a|X^n), \tag{13.75}$$

and the expectation of the indicator random variable $I(X = a)$ is equal to the probability of the symbol a:

$$\mathbb{E}_X\{I(X = a)\} = p_X(a). \tag{13.76}$$

Also, any random sequence X^n has probability zero if one of its symbols x_i is such that $p_X(x_i) = 0$. Thus, the probability that $\frac{1}{n} N(a|X^n) = 0$ is equal to one whenever $p_X(a) = 0$:

$$\Pr\left\{ \frac{1}{n} N(a|X^n) = 0 : p_X(a) = 0 \right\} = 1, \tag{13.77}$$

and we can consider the cases when $p_X(a) > 0$. We apply the law of large numbers to find that

$$\forall \epsilon, \delta > 0 \;\; \exists n_{0,a} : \forall n > n_{0,a} \qquad \Pr\left\{ \left| \frac{1}{n} N(a|X^n) - p_X(a) \right| > \delta \right\} < \frac{\epsilon}{|\mathcal{X}|}. \tag{13.78}$$

Choosing $n_0 = \max_{a \in \mathcal{X}} \{n_{0,a}\}$, the following condition holds by the union bound of probability theory:

$$\begin{aligned} \forall \epsilon, \delta > 0 \;\; \exists n_0 : \forall n > n_0 & \\ \Pr\left\{ \bigcup_{a \in \mathcal{X}} \left| \frac{1}{n} N(a|X^n) - p_X(a) \right| > \delta \right\} & \\ \leq \sum_{a \in \mathcal{X}} \Pr\left\{ \left| \frac{1}{n} N(a|X^n) - p_X(a) \right| > \delta \right\} < \epsilon. & \end{aligned} \tag{13.79}$$

Thus it holds that the complement of the above event on the left holds with high probability:

$$\forall \epsilon, \delta > 0 \;\; \exists n_0 : \forall n > n_0 \qquad \Pr\left\{ \forall a \in \mathcal{X}, \; \left| \frac{1}{n} N(a|X^n) - p_X(a) \right| \leq \delta \right\} > 1 - \epsilon. \tag{13.80}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The event $\left\{\forall a \in \mathcal{X},\ \left|\frac{1}{n}N(a|X^n) - p_X(a)\right| \leq \delta\right\}$ is the condition for a random sequence X^n to be in the strongly typical set $T_\delta^{X^n}$, and the probability of this event goes to one as n becomes sufficiently large. □

Proof of the Exponentially Small Cardinality Property (Property 13.7.3). By the proof of Property 13.7.4 (proved below), we know that the following relation holds for any sequence x^n in the strongly typical set:

$$2^{-n(H(X)+c\delta)} \leq p_{X^n}(x^n) \leq 2^{-n(H(X)-c\delta)}, \tag{13.81}$$

where c is some constant that we define when we prove Property 13.7.4. Summing over all sequences in the typical set, we get the following inequalities:

$$\sum_{x^n \in T_\delta^{X^n}} 2^{-n(H(X)+c\delta)} \leq \Pr\left\{X^n \in T_\delta^{X^n}\right\} \leq \sum_{x^n \in T_\delta^{X^n}} 2^{-n(H(X)-c\delta)}, \tag{13.82}$$

$$\Rightarrow 2^{-n(H(X)+c\delta)}\left|T_\delta^{X^n}\right| \leq \Pr\left\{X^n \in T_\delta^{X^n}\right\} \leq 2^{-n(H(X)-c\delta)}\left|T_\delta^{X^n}\right|. \tag{13.83}$$

By the unit probability property of the strongly typical set, we know that the following relation holds for sufficiently large n:

$$1 \geq \Pr\left\{X^n \in T_\delta^{X^n}\right\} \geq 1 - \epsilon. \tag{13.84}$$

Then the following inequalities result by combining the above inequalities:

$$2^{n(H(X)-c\delta)}(1-\epsilon) \leq \left|T_\delta^{X^n}\right| \leq 2^{n(H(X)+c\delta)}. \tag{13.85}$$

□

Proof of the Equipartition Property (Property 13.7.4). The following relation holds from the IID property of the distribution $p_{X^n}(x^n)$ and because the sequence x^n is strongly typical:

$$p_{X^n}(x^n) = \prod_{x \in \mathcal{X}^+} p_X(x)^{N(x|x^n)}, \tag{13.86}$$

where $\mathcal{X}^+$ denotes all the letters x in $\mathcal{X}$ with $p_X(x) > 0$. (The fact that the sequence x^n is strongly typical according to Definition 13.7.2 allows us to employ this modified alphabet). Take the logarithm of the above expression:

$$\log(p_{X^n}(x^n)) = \sum_{x \in \mathcal{X}^+} N(x|x^n)\log(p_X(x)), \tag{13.87}$$

Multiply both sides by $-\frac{1}{n}$:

$$-\frac{1}{n}\log(p_{X^n}(x^n)) = -\sum_{x \in \mathcal{X}^+} \frac{1}{n}N(x|x^n)\log(p_X(x)), \tag{13.88}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The following relation holds because the sequence x^n is strongly typical:

$$\forall x \in \mathcal{X}^+ : \left| \frac{1}{n} N(x|x^n) - p_X(x) \right| \leq \delta, \tag{13.89}$$

and it implies that

$$\Rightarrow \forall x \in \mathcal{X}^+ : -\delta + p_X(x) \leq \frac{1}{n} N(x|x^n) \leq \delta + p_X(x). \tag{13.90}$$

Now multiply (13.90) by $-\log(p_X(x)) > 0$, sum over all letters in the alphabet $\mathcal{X}^+$, and apply the substitution in (13.88). This procedure gives the following set of inequalities:

$$\begin{aligned} -\sum_{x \in \mathcal{X}^+} (-\delta + p_X(x)) \log(p_X(x)) &\leq -\frac{1}{n} \log(p_{X^n}(x^n)) \\ &\leq -\sum_{x \in \mathcal{X}^+} (\delta + p_X(x)) \log(p_X(x)), \end{aligned} \tag{13.91}$$

$$\Rightarrow -c\delta + H(X) \leq -\frac{1}{n} \log(p_{X^n}(x^n)) \leq c\delta + H(X), \tag{13.92}$$

$$\Rightarrow 2^{-n(H(X)+c\delta)} \leq p_{X^n}(x^n) \leq 2^{-n(H(X)-c\delta)}, \tag{13.93}$$

where

$$c \equiv -\sum_{x \in \mathcal{X}^+} \log(p_X(x)) \geq 0. \tag{13.94}$$

It now becomes apparent why we require the technical condition in the definition of strong typicality (Definition 13.7.2). Were it not there, then the constant c would not be finite, and we would not be able to obtain a reasonable bound on the probability of a strongly typical sequence. □

13.7.4 Cardinality of a Typical Type Class

Recall that a typical type class is the set of all sequences with the same empirical distribution, and the empirical distribution happens to have maximum deviation δ from the true distribution. It might seem that the size $\left|T_t^{X^n}\right|$ of a typical type class $T_t^{X^n}$ should be smaller than the size of the strongly typical set. But the following property overrides this intuition and shows that a given typical type class $T_t^{X^n}$ has almost as many sequences in it as the strongly typical set $T_\delta^{X^n}$ for sufficiently large n.

Property 13.7.5 (Minimal Cardinality of a Typical Type Class) For $t \in \tau_\delta$ and for sufficiently large n, the size $\left|T_t^{X^n}\right|$ of the typical type class $T_t^{X^n}$ is lower bounded as follows:

$$\left|T_t^{X^n}\right| \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{n[H(X)-\eta(|\mathcal{X}|\delta)]} = 2^{n\left[H(X)-\eta(|\mathcal{X}|\delta)-|\mathcal{X}|\frac{1}{n}\log(n+1)\right]}, \tag{13.95}$$

where $\eta(\delta)$ is some function such that $\eta(\delta) \to 0$ as $\delta \to 0$. Thus, a typical type class is of size roughly $2^{nH(X)}$ when $n \to \infty$ and $\delta \to 0$ (it is about as large as the typical set when n becomes large).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Proof. We first show that if $X_1, \ldots, X_n$ are random variables drawn IID from a distribution $q(x)$, then the probability $q^n(x^n)$ of a particular sequence x^n depends only on its type:

$$q^n(x^n) = 2^{-n(H(t_{x^n})+D(t_{x^n}||q))}, \tag{13.96}$$

where $D(t_{x^n}||q)$ is the relative entropy between t_{x^n} and q. Consider the following chain of equalities:

$$\begin{aligned} q^n(x^n) &= \prod_{i=1}^{n} q(x_i) = \prod_{x\in\mathcal{X}} q(x)^{N(x|x^n)} = \prod_{x\in\mathcal{X}} q(x)^{nt_{x^n}(x)} && (13.97)\\ &= \prod_{x\in\mathcal{X}} 2^{nt_{x^n}(x)\log q(x)} = 2^{n\sum_{x\in\mathcal{X}} t_{x^n}(x)\log q(x)} && (13.98)\\ &= 2^{n\sum_{x\in\mathcal{X}} t_{x^n}(x)\log q(x) - t_{x^n}(x)\log t_{x^n}(x) + t_{x^n}(x)\log t_{x^n}(x)} && (13.99)\\ &= 2^{-n(D(t_{x^n}||q)+H(t_{x^n}))} && (13.100) \end{aligned}$$

It then follows that the probability of the sequence x^n is $2^{-nH(t_{x^n})}$ if the distribution $q(x) = t_{x^n}(x)$. Now consider that each type class $T_t^{X^n}$ has size

$$\binom{n}{nt_{x^n}(x_1), \ldots, nt_{x^n}\big(x_{|\mathcal{X}|}\big)}, \tag{13.101}$$

where the distribution $t = \big(t_{x^n}(x_1), \ldots, t_{x^n}\big(x_{|\mathcal{X}|}\big)\big)$ and the letters of $\mathcal{X}$ are $x_1, \ldots, x_{|\mathcal{X}|}$. This result follows because the size of a type class is just the number of ways of arranging $nt_{x^n}(x_1), \ldots, nt_{x^n}\big(x_{|\mathcal{X}|}\big)$ in a sequence of length n. We now prove that the type class $T_t^{X^n}$ has the highest probability among all type classes when the probability distribution is t:

$$t^n\big(T_t^{X^n}\big) \geq t^n\big(T_{t'}^{X^n}\big) \text{ for all } t' \in \mathcal{P}_n, \tag{13.102}$$

where t^n is the IID distribution induced by the type t and $\mathcal{P}_n$ is the set of all types. Consider the following equalities:

$$\begin{aligned} \frac{t^n\big(T_t^{X^n}\big)}{t^n\big(T_{t'}^{X^n}\big)} &= \frac{\big|T_t^{X^n}\big| \prod_{x\in\mathcal{X}} t_{x^n}(x)^{nt_{x^n}(x)}}{\big|T_{t'}^{X^n}\big| \prod_{x\in\mathcal{X}} t_{x^n}(x)^{nt'_{x^n}(x)}} && (13.103)\\ &= \frac{\binom{n}{nt_{x^n}(x_1),\ldots,nt_{x^n}\left(x_{|\mathcal{X}|}\right)} \prod_{x\in\mathcal{X}} t_{x^n}(x)^{nt_{x^n}(x)}}{\binom{n}{nt'_{x^n}(x_1),\ldots,nt'_{x^n}\left(x_{|\mathcal{X}|}\right)} \prod_{x\in\mathcal{X}} t_{x^n}(x)^{nt'_{x^n}(x)}} && (13.104)\\ &= \prod_{x\in\mathcal{X}} \frac{nt'_{x^n}(x)!}{nt_{x^n}(x)!} t_{x^n}(x)^{n\left(t_{x^n}(x)-t'_{x^n}(x)\right)}. && (13.105) \end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Now apply the bound $\frac{m!}{n!} \geq n^{m-n}$ (that holds for any positive integers m and n) to get

$$\frac{t^n\left(T_t^{X^n}\right)}{t^n\left(T_{t'}^{X^n}\right)} \geq \prod_{x\in\mathcal{X}} [nt_{x^n}(x)]^{n\left(t'_{x^n}(x)-t_{x^n}(x)\right)} t_{x^n}(x)^{n\left(t_{x^n}(x)-t'_{x^n}(x)\right)} \tag{13.106}$$

$$= \prod_{x\in\mathcal{X}} n^{n\left(t'_{x^n}(x)-t_{x^n}(x)\right)} \tag{13.107}$$

$$= n^{n\sum_{x\in\mathcal{X}} t'_{x^n}(x)-t_{x^n}(x)} \tag{13.108}$$

$$= n^{n(1-1)} \tag{13.109}$$

$$= 1. \tag{13.110}$$

Thus, it holds that $t^n\left(T_t^{X^n}\right) \geq t^n\left(T_{t'}^{X^n}\right)$ for all t'. Now we are close to obtaining the desired bound in Property 13.7.5. Consider the following chain of inequalities:

$$1 = \sum_{t'\in\mathcal{P}_n} t^n\left(T_{t'}^{X^n}\right) \leq \sum_{t'\in\mathcal{P}_n} \max_{t'} t^n\left(T_{t'}^{X^n}\right) = \sum_{t'\in\mathcal{P}_n} t^n\left(T_t^{X^n}\right) \tag{13.111}$$

$$\leq (n+1)^{|\mathcal{X}|} t^n\left(T_t^{X^n}\right) = (n+1)^{|\mathcal{X}|} \sum_{x^n\in T_t^{X^n}} t^n(x^n) \tag{13.112}$$

$$= (n+1)^{|\mathcal{X}|} \sum_{x^n\in T_t^{X^n}} 2^{-nH(t)} \tag{13.113}$$

$$= (n+1)^{|\mathcal{X}|} 2^{-nH(t)} \left|T_t^{X^n}\right| \tag{13.114}$$

Recall that t is a typical type, implying that $|t(x) - p(x)| \leq \delta$ for all x. This then implies that the variational distance between the distributions is small:

$$\sum_x |t(x) - p(x)| \leq |\mathcal{X}|\delta. \tag{13.115}$$

We can apply Fannes' inequality for continuity of entropy (Theorem 11.9.5) to get a bound on the difference of entropies:

$$|H(t) - H(X)| \leq 2|\mathcal{X}|\delta \log|\mathcal{X}| + 2H_2(|\mathcal{X}|\delta). \tag{13.116}$$

The desired bound then follows with $\eta(|\mathcal{X}|\delta) \equiv 2|\mathcal{X}|\delta \log|\mathcal{X}| + 2H_2(|\mathcal{X}|\delta)$. □

Exercise 13.7.1 Prove that $2^{nH(t)}$ is an upper bound on the number of sequences x^n of type t:

$$\left|T_t^{X^n}\right| \leq 2^{nH(t)}. \tag{13.117}$$

Use this bound and (13.100) to prove the following upper bound on the probability of a type class where each sequence is generated IID according to a distribution $q(x)$:

$$\Pr\left\{T_t^{X^n}\right\} \leq 2^{-nD(t \,\|\, q)}. \tag{13.118}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

13.8 Strong Joint Typicality

It is possible to extend the above notions of strong typicality to jointly typical sequences. In a marked difference with the weakly typical case, we can show that strong joint typicality implies marginal typicality. Thus there is no need to impose this constraint in the definition.

Let $N(x, y|x^n, y^n)$ be the number of occurrences of the symbol $x \in \mathcal{X}, y \in \mathcal{Y}$ in the respective sequences x^n and y^n. The *type* or empirical distribution $t_{x^n y^n}$ of sequences x^n and y^n is a probability mass function whose elements are $t_{x^n y^n}(x, y)$ where

$$t_{x^n y^n}(x, y) \equiv \frac{1}{n} N(x, y|x^n, y^n). \tag{13.119}$$

Definition 13.8.1 (Strong Jointly Typical Sequence)**.** *Two sequences x^n, y^n are δ-strongly-jointly-typical if their empirical distribution has maximum deviation δ from the true distribution and vanishes for any two symbols x and y for which $p_{X,Y}(x, y) = 0$.*

Definition 13.8.2 (Strong Jointly Typical Set)**.** *The δ-jointly typical set $T_\delta^{X^n Y^n}$ is the set of all δ-jointly typical sequences:*

$$T_\delta^{X^n Y^n} \equiv \left\{ x^n, y^n : \forall (x, y) \in \mathcal{X} \times \mathcal{Y} \begin{array}{cc} \left|\frac{1}{n} N(x, y|x^n, y^n) - p_{X,Y}(x, y)\right| \leq \delta & \text{if } p_{X,Y}(x, y) > 0 \\ \frac{1}{n} N(x, y|x^n, y^n) = 0 & \text{otherwise} \end{array} \right\}. \tag{13.120}$$

It follows from the above definitions that strong joint typicality implies marginal typicality for both sequences x^n and y^n. We leave the proof as the following exercise.

Exercise 13.8.1 Prove that strong joint typicality implies marginal typicality for either the sequence x^n or y^n.

13.8.1 Properties of the Strong Jointly Typical Set

The set $T_\delta^{X^n Y^n}$ of strong jointly typical sequences enjoys properties similar to what we have seen before.

Property 13.8.1 (Unit Probability) The strong jointly typical set $T_\delta^{X^n Y^n}$ asymptotically has probability one. So as n becomes large, it is highly likely that a source emits a strong jointly typical sequence. We formally state this property as follows:

$$\forall \epsilon > 0 \qquad \Pr\left\{ X^n Y^n \in T_\delta^{X^n Y^n} \right\} \geq 1 - \epsilon \quad \text{for sufficiently large } n. \tag{13.121}$$

Property 13.8.2 (Exponentially Small Cardinality) The number $\left|T_\delta^{X^n Y^n}\right|$ of δ-jointly typical sequences is exponentially smaller than the total number $(|\mathcal{X}||\mathcal{Y}|)^n$ of sequences for any joint random variable (X, Y) that is not uniform. We formally state this property as follows:

$$\left|T_\delta^{X^n Y^n}\right| \leq 2^{n(H(X,Y)+c\delta)}, \tag{13.122}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where c is a constant. We can also lower bound the size of the δ-jointly typical set when n is sufficiently large:

$$\forall \epsilon > 0 \qquad \left|T_\delta^{X^n Y^n}\right| \geq (1-\epsilon)2^{n(H(X,Y)-c\delta)} \qquad \text{for sufficiently large } n. \tag{13.123}$$

Property 13.8.3 (Equipartition) The probability of a given δ-jointly typical sequence $x^n y^n$ occurring is approximately uniform:

$$2^{-n(H(X,Y)+c\delta)} \leq p_{X^n,Y^n}(x^n, y^n) \leq 2^{-n(H(X,Y)-c\delta)}. \tag{13.124}$$

Property 13.8.4 (Probability of Strong Joint Typicality) Consider two independent random variables $\tilde{X}^n$ and $\tilde{Y}^n$ whose respective probability density functions $p_{\tilde{X}^n}(x^n)$ and $p_{\tilde{Y}^n}(y^n)$ are equal to the marginal densities of the joint density $p_{X^n,Y^n}(x^n, y^n)$:

$$\left(\tilde{X}^n, \tilde{Y}^n\right) \sim p_{X^n}(x^n) p_{Y^n}(y^n). \tag{13.125}$$

Then we can bound the probability that two random sequences $\tilde{X}^n$ and $\tilde{Y}^n$ are in the jointly typical set $T_\delta^{X^n Y^n}$:

$$\Pr\left\{\left(\tilde{X}^n, \tilde{Y}^n\right) \in T_\delta^{X^n Y^n}\right\} \leq 2^{-n(I(X;Y)-3c\delta)}. \tag{13.126}$$

The proofs of the first three properties are the same as in the previous section, and the proof of the last property is the same as that for the weakly typical case.

13.9 Strong Conditional Typicality

Strong conditional typicality bears some similarities to strong typicality, but it is sufficiently different for us to provide a discussion of it. We first introduce it with a simple example.

Suppose that we draw a sequence from an alphabet $\{0, 1, 2\}$ according to the distribution:

$$p_X(0) = \frac{1}{4}, \qquad p_X(1) = \frac{1}{4}, \qquad p_X(2) = \frac{1}{2}. \tag{13.127}$$

A particular realization sequence could be as follows:

$$201020102012021212222020222 2. \tag{13.128}$$

We count up the occurrences of each symbol and find them to be

$$N(0 \mid 201020102012021212222020222 2) = 8, \tag{13.129}$$
$$N(1 \mid 201020102012021212222020222 2) = 5, \tag{13.130}$$
$$N(2 \mid 201020102012021212222020222 2) = 15. \tag{13.131}$$

The maximum deviation of the sequence's empirical distribution from the true distribution of the source is as follows:

$$\max\left\{\left|\frac{1}{4} - \frac{8}{28}\right|, \left|\frac{1}{4} - \frac{5}{28}\right|, \left|\frac{1}{2} - \frac{15}{28}\right|\right\} = \max\left\{\frac{1}{28}, \frac{2}{28}, \frac{1}{28}\right\} = \frac{1}{14}. \tag{13.132}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We now consider generating a different sequence from an alphabet $\{a, b, c\}$. Though, we generate it according to the following *conditional* probability distribution:

$$\begin{bmatrix} p_{Y|X}(a|0) = \frac{1}{5} & p_{Y|X}(a|1) = \frac{1}{6} & p_{Y|X}(a|2) = \frac{2}{4} \\ p_{Y|X}(b|0) = \frac{2}{5} & p_{Y|X}(b|1) = \frac{3}{6} & p_{Y|X}(b|2) = \frac{1}{4} \\ p_{Y|X}(c|0) = \frac{2}{5} & p_{Y|X}(c|1) = \frac{2}{6} & p_{Y|X}(c|2) = \frac{1}{4} \end{bmatrix}. \tag{13.133}$$

The second generated sequence should thus have correlations with the original sequence. A possible realization of the second sequence could be as follows:

$$abbcbccabcabcabcabcbcbabacba. \tag{13.134}$$

We would now like to analyze how close the empirical conditional distribution is to the true conditional distribution for all input and output sequences. A useful conceptual first step is to apply a permutation to the first sequence so that all of its symbols appear in lexicographic order, and we then apply the same permutation to the second sequence:

$$\begin{array}{cccccccccccccccccccccccc} 2 & 0 & 1 & 0 & 2 & 0 & 1 & 0 & 2 & 0 & 1 & 2 & 0 & 2 & 1 & 2 & 1 & 2 & 2 & 2 & 2 & 0 & 2 & 0 & 2 & 2 & 2 & 2 \\ a & b & b & c & b & c & c & a & b & c & a & b & c & a & b & c & a & b & c & b & c & b & a & b & a & c & b & a \end{array}$$

$$\xrightarrow{\text{permute}}$$

$$\begin{array}{cccccccccccccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ b & c & c & a & c & c & b & b & b & c & a & b & a & a & b & b & b & a & c & b & c & b & c & a & a & c & b & a \end{array}. \tag{13.135}$$

This rearrangement makes it easy to count up the empirical conditional distribution of the second sequence. We first place the joint occurrences of the symbols into the following matrix:

$$\begin{bmatrix} N(0,a) = 1 & N(1,a) = 2 & N(2,a) = 5 \\ N(0,b) = 3 & N(1,b) = 2 & N(2,b) = 6 \\ N(0,c) = 4 & N(1,c) = 1 & N(2,c) = 4 \end{bmatrix}, \tag{13.136}$$

and we obtain the empirical conditional distribution matrix by dividing these entries by the marginal distribution of the first sequence:

$$\begin{bmatrix} \frac{N(0,a)}{N(0)} = \frac{1}{8} & \frac{N(1,a)}{N(1)} = \frac{2}{5} & \frac{N(2,a)}{N(2)} = \frac{5}{15} \\ \frac{N(0,b)}{N(0)} = \frac{3}{8} & \frac{N(1,b)}{N(1)} = \frac{2}{5} & \frac{N(2,b)}{N(2)} = \frac{6}{15} \\ \frac{N(0,c)}{N(0)} = \frac{4}{8} & \frac{N(1,c)}{N(1)} = \frac{1}{5} & \frac{N(2,c)}{N(2)} = \frac{4}{15} \end{bmatrix}. \tag{13.137}$$

We then compare the maximal deviation of the elements in this matrix with the elements in the stochastic matrix in (13.133):

$$\begin{aligned} &\max\left\{\left|\frac{1}{5}-\frac{1}{8}\right|, \left|\frac{2}{5}-\frac{3}{8}\right|, \left|\frac{2}{5}-\frac{4}{8}\right|, \left|\frac{1}{6}-\frac{2}{5}\right|, \left|\frac{3}{6}-\frac{2}{5}\right|, \left|\frac{2}{6}-\frac{1}{5}\right|, \left|\frac{2}{4}-\frac{5}{15}\right|, \left|\frac{1}{4}-\frac{6}{15}\right|, \left|\frac{1}{4}-\frac{4}{15}\right|\right\} \\ &\qquad = \max\left\{\frac{3}{40}, \frac{1}{40}, \frac{1}{10}, \frac{7}{30}, \frac{1}{10}, \frac{2}{15}, \frac{1}{6}, \frac{3}{20}, \frac{1}{60}\right\} = \frac{7}{30}. \end{aligned} \tag{13.138}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above analysis applies to a finite realization to illustrate the notion of conditional typicality, and there is a large deviation from the true distribution in this case. We would again expect this deviation to vanish for a random sequence in the limit as the length of the sequence becomes asymptotically large.

13.9.1 Definition of Strong Conditional Typicality

We now give a formal definition of strong conditional typicality.

Definition 13.9.1 (Conditional Empirical Distribution)**.** *The conditional empirical distribution $t_{y^n|x^n}(y|x)$ is as follows:*

$$t_{y^n|x^n}(y|x) = \frac{t_{x^n y^n}(x,y)}{t_{x^n}(x)}. \tag{13.139}$$

Definition 13.9.2 (Strong Conditional Typicality)**.** *Suppose that a sequence x^n is a strongly typical sequence in $T_\delta^{X^n}$. Then the δ-strong conditionally typical set $T_\delta^{Y^n|x^n}$ corresponding to the sequence x^n consists of all sequences whose joint empirical distribution $\frac{1}{n}N(x,y|x^n,y^n)$ is δ-close to the product of the true conditional distribution $p_{Y|X}(y|x)$ with the marginal empirical distribution $\frac{1}{n}N(x|x^n)$:*

$$T_\delta^{Y^n|x^n} \equiv$$
$$\left\{ y^n : \forall (x,y) \in \mathcal{X} \times \mathcal{Y} \begin{array}{cc} |N(x,y|x^n,y^n) - p(y|x)N(x|x^n)| \leq n\delta & \text{if } p(y|x) > 0 \\ N(x,y|x^n,y^n) = 0 & \text{otherwise} \end{array} \right\}, \tag{13.140}$$

where we abbreviate $p_{Y|X}(y|x)$ as $p(y|x)$

The above definition of strong conditional typicality implies that the conditional empirical distribution is close to the true conditional distribution, in the sense that

$$\left| \frac{t_{x^n y^n}(x,y)}{t_{x^n}(x)} - p_{Y|X}(y|x) \right| \leq \frac{1}{t_{x^n}(x)}\delta. \tag{13.141}$$

Of course, such a relation only makes sense if the marginal empirical distribution $t_{x^n}(x)$ is non-zero.

The extra technical condition ($N(x,y|x^n,y^n) = 0$ if $p_{Y|X}(y|x) = 0$) in Definition 13.9.2 is present again for a reason that we found in the proof of the Equipartition Property for Strong Typicality (Property 13.7.4).

13.9.2 Properties of the Strong Conditionally Typical Set

The set $T_\delta^{Y^n|x^n}$ of conditionally typical sequences enjoys a few useful properties that are similar to what we have for the weak conditionally typical set, but the initial sequence x^n can be deterministic. Though, we do impose the constraint that it has to be strongly typical so that we can prove useful properties for the corresponding strong conditionally typical set. So first suppose that a given sequence $x^n \in T_{\delta'}^{X^n}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Property 13.9.1 (Unit Probability) The set $T_\delta^{Y^n|x^n}$ asymptotically has probability one. So as n becomes large, it is highly likely that a random sequence Y^n corresponding to a given typical sequence x^n is a conditionally typical sequence. We formally state this property as follows:

$$\forall \epsilon > 0 \qquad \Pr\left\{ Y^n \in T_\delta^{Y^n|x^n} \right\} \geq 1 - \epsilon \quad \text{for sufficiently large } n. \tag{13.142}$$

Property 13.9.2 (Exponentially Small Cardinality) The number $\left| T_\delta^{Y^n|x^n} \right|$ of δ-conditionally typical sequences is exponentially smaller than the total number $|\mathcal{Y}|^n$ of sequences for any conditional random variable Y that is not uniform. We formally state this property as follows:

$$\left| T_\delta^{Y^n|x^n} \right| \leq 2^{n(H(Y|X)+c(\delta+\delta'))}. \tag{13.143}$$

We can also lower bound the size of the δ-conditionally typical set when n is sufficiently large:

$$\forall \epsilon > 0 \qquad \left| T_\delta^{Y^n|x^n} \right| \geq (1-\epsilon) 2^{n(H(Y|X)-c(\delta+\delta'))} \quad \text{for sufficiently large } n. \tag{13.144}$$

Property 13.9.3 (Equipartition) The probability of a particular δ-conditionally typical sequence y^n is approximately uniform:

$$2^{-n(H(Y|X)+c(\delta+\delta'))} \leq p_{Y^n|X^n}(y^n|x^n) \leq 2^{-n(H(Y|X)-c(\delta+\delta'))}. \tag{13.145}$$

In summary, given a realization x^n of the random variable X^n, the conditionally typical set $T_\delta^{Y^n|x^n}$ has almost all the probability, its size is exponentially smaller than the size of the set of all sequences, and each δ-conditionally typical sequence has an approximate uniform probability of occurring.

13.9.3 Proofs of the Properties of the Strong Conditionally Typical Set

Proof of the Unit Probability Property (Property 13.9.1. The proof of this property is somewhat more complicated for strong conditional typicality. Since we are dealing with an IID distribution, we can assume that the sequence x^n is lexicographically ordered with an order on the alphabet $\mathcal{X}$. We write the elements of $\mathcal{X}$ as $x_1, \ldots, x_{|\mathcal{X}|}$. Then the lexicographic ordering means that we can write the sequence x^n as follows:

$$x^n = \underbrace{x_1 \cdots x_1}_{N(x_1|x^n)} \underbrace{x_2 \cdots x_2}_{N(x_2|x^n)} \cdots \underbrace{x_{|\mathcal{X}|} \cdots x_{|\mathcal{X}|}}_{N\left(x_{|\mathcal{X}|}|x^n\right)}. \tag{13.146}$$

It follows that $N(x|x^n) \geq n(p_X(x) - \delta')$ from the typicality of x^n, and the law of large numbers comes into play for each block $x_i \cdots x_i$ with length $N(x_i|x^n)$ when this length is large

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

enough. Let $p_{Y|X=x}(y)$ be the distribution for the conditional random variable $Y|(X=x)$. Then the following set is a slightly stronger notion of conditional typicality:

$$\left\{y^n \in T_\delta^{Y^n|x^n}\right\} \Leftrightarrow \bigwedge_{x\in\mathcal{X}} \left\{y^{N(x|x^n)} \in T_\delta^{(Y|(X=x))^{N(x|x^n)}}\right\}, \tag{13.147}$$

where the symbol $\wedge$ denotes concatenation (note that the lexicographic ordering of x^n applies to the ordering of the sequence y^n as well). Also, $T_\delta^{(Y|(X=x))^{N(x|x^n)}}$ is the typical set for a sequence of conditional random variables $Y|(X=x)$ with length $N(x|x^n)$:

$$T_\delta^{(Y|(X=x))^{N(x|x^n)}} \equiv \left\{y^{N(x|x^n)} : \forall y \in \mathcal{Y}, \quad \left|\frac{N\big(y|y^{N(x|x^n)}\big)}{N(x|x^n)} - p_{Y|X=x}(y)\right| \leq \delta\right\}. \tag{13.148}$$

We can apply the law of large numbers to each of these typical sets $T_\delta^{(Y|(X=x))^{N(x|x^n)}}$ where the length $N(x|x^n)$ becomes large. It then follows that

$$\begin{aligned}
\Pr\left\{Y^n \in T_\delta^{Y^n|x^n}\right\} &= \prod_{x\in\mathcal{X}} \Pr\left\{Y^{N(x|x^n)} \in T_\delta^{(Y|(X=x))^{N(x|x^n)}}\right\} && (13.149)\\
&\geq (1-\epsilon)^{|\mathcal{X}|} && (13.150)\\
&\geq 1 - |\mathcal{X}|\epsilon. && (13.151)
\end{aligned}$$

□

Proof of the Equipartition Property (Property 13.9.3). The following relation holds from the IID property of the conditional distribution $p_{Y^n|X^n}(y^n|x^n)$ and because the sequence y^n is strong conditionally typical according to Definition 13.9.2:

$$p_{Y^n|X^n}(y^n|x^n) = \prod_{(\mathcal{X},\mathcal{Y})^+} p_{Y|X}(y|x)^{N(x,y|x^n,y^n)}, \tag{13.152}$$

where $(\mathcal{X},\mathcal{Y})^+$ denotes all the letters x, y in $\mathcal{X}, \mathcal{Y}$ with $p_{Y|X}(y|x) > 0$. Take the logarithm of the above expression:

$$\log\big(p_{Y^n|X^n}(y^n|x^n)\big) = \sum_{x,y\in(\mathcal{X},\mathcal{Y})^+} N(x,y|x^n,y^n)\log\big(p_{Y|X}(y|x)\big), \tag{13.153}$$

Multiply both sides by $-\frac{1}{n}$:

$$-\frac{1}{n}\log\big(p_{Y^n|X^n}(y^n|x^n)\big) = -\sum_{x,y\in(\mathcal{X},\mathcal{Y})^+} \frac{1}{n} N(x,y|x^n,y^n)\log\big(p_{Y|X}(y|x)\big). \tag{13.154}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The following relations hold because the sequence x^n is strongly typical and y^n is strong conditionally typical:

$$\forall x \in \mathcal{X}^+ : \left|\frac{1}{n}N(x|x^n) - p_X(x)\right| \leq \delta', \tag{13.155}$$

$$\Rightarrow \forall x \in \mathcal{X}^+ : -\delta' + p_X(x) \leq \frac{1}{n}N(x|x^n) \leq \delta' + p_X(x), \tag{13.156}$$

$$\forall x, y \in (\mathcal{X},\mathcal{Y})^+ : \left|\frac{1}{n}N(x,y|x^n,y^n) - p_{Y|X}(y|x)\frac{1}{n}N(x|x^n)\right| \leq \delta \tag{13.157}$$

$$\begin{aligned}\Rightarrow \forall x, y \in (\mathcal{X},\mathcal{Y})^+ : -\delta + p_{Y|X}(y|x)\frac{1}{n}N(x|x^n) &\leq \frac{1}{n}N(x,y|x^n,y^n) \\ &\leq \delta + p_{Y|X}(y|x)\frac{1}{n}N(x|x^n)\end{aligned} \tag{13.158}$$

Now multiply (13.158) by $-\log\big(p_{Y|X}(y|x)\big) > 0$, sum over all letters in the alphabet $(\mathcal{X},\mathcal{Y})^+$, and apply the substitution in (13.154). This procedure gives the following set of inequalities:

$$\begin{aligned}-\sum_{x,y\in(\mathcal{X},\mathcal{Y})^+}\left(-\delta + p_{Y|X}(y|x)\frac{1}{n}N(x|x^n)\right)\log\big(p_{Y|X}(y|x)\big) & \\ \leq -\frac{1}{n}\log\big(p_{Y^n|X^n}(y^n|x^n)\big) & \\ \leq -\sum_{x,y\in(\mathcal{X},\mathcal{Y})^+}\left(\delta + p_{Y|X}(y|x)\frac{1}{n}N(x|x^n)\right)\log\big(p_{Y|X}(y|x)\big), &\end{aligned} \tag{13.159}$$

Now apply the inequalities in (13.156) (assuming that $p_X(x) \geq \delta'$ for $x \in \mathcal{X}^+$) to get that

$$\Rightarrow -\sum_{x,y\in(\mathcal{X},\mathcal{Y})^+}\big(-\delta + p_{Y|X}(y|x)(-\delta' + p_X(x))\big)\log\big(p_{Y|X}(y|x)\big) \tag{13.160}$$

$$\leq -\frac{1}{n}\log\big(p_{Y^n|X^n}(y^n|x^n)\big) \tag{13.161}$$

$$\leq -\sum_{x,y\in(\mathcal{X},\mathcal{Y})^+}\big(\delta + p_{Y|X}(y|x)(\delta' + p_X(x))\big)\log\big(p_{Y|X}(y|x)\big) \tag{13.162}$$

$$\Rightarrow -c(\delta+\delta') + H(Y|X) \leq -\frac{1}{n}\log\big(p_{Y^n|X^n}(y^n|x^n)\big) \leq c(\delta+\delta') + H(Y|X), \tag{13.163}$$

$$\Rightarrow 2^{-n(H(Y|X)+c(\delta+\delta'))} \leq p_{Y^n|X^n}(y^n|x^n) \leq 2^{-n(H(Y|X)-c(\delta+\delta'))}, \tag{13.164}$$

where

$$c \equiv -\sum_{x,y\in(\mathcal{X},\mathcal{Y})^+}\log\big(p_{Y|X}(y|x)\big) \geq 0. \tag{13.165}$$

It again becomes apparent why we require the technical condition in the definition of strong conditional typicality (Definition 13.9.2). Were it not there, then the constant c would not

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

be finite, and we would not be able to obtain a reasonable bound on the probability of a strong conditionally typical sequence. □

We close this section with a lemma that relates strong conditional, marginal, and joint typicality.

Lemma 13.9.1. *Suppose that y^n is a conditionally typical sequence in $T_{\delta}^{Y^n|x^n}$ and its conditioning sequence x^n is a typical sequence in $T_{\delta'}^{X^n}$. Then x^n and y^n are jointly typical in the set $T_{\delta+\delta'}^{X^nY^n}$, and y^n is a typical sequence in $T_{|\mathcal{X}|(\delta+\delta')}^{Y^n}$.*

Proof. It follows from the above that $\forall x \in \mathcal{X}, y \in \mathcal{Y}$:

$$p_X(x) - \delta' \leq \frac{1}{n}N(x|x^n) \leq \delta' + p_X(x), \tag{13.166}$$

$$p_{Y|X}(y|x)\frac{1}{n}N(x|x^n) - \delta \leq \frac{1}{n}N(x,y|x^ny^n) \leq \delta + p_{Y|X}(y|x)\frac{1}{n}N(x|x^n). \tag{13.167}$$

Substituting the upper bound on $\frac{1}{n}N(x|x^n)$ gives

$$\begin{aligned}\frac{1}{n}N(x,y|x^ny^n) &\leq \delta + p_{Y|X}(y|x)(\delta' + p_X(x)) &(13.168)\\ &= \delta + p_{Y|X}(y|x)\delta' + p_X(x)p_{Y|X}(y|x) &(13.169)\\ &\leq \delta + \delta' + p_{X,Y}(x,y). &(13.170)\end{aligned}$$

Similarly, substituting the lower bound on $\frac{1}{n}N(x|x^n)$ gives

$$\frac{1}{n}N(x,y|x^ny^n) \geq p_{X,Y}(x,y) - \delta - \delta'. \tag{13.171}$$

Putting both of the above bounds together, we get the following bound:

$$\left|\frac{1}{n}N(x,y|x^ny^n) - p_{X,Y}(x,y)\right| \leq \delta + \delta'. \tag{13.172}$$

This then implies that the sequences x^n and y^n lie in the strong jointly typical set $T_{\delta+\delta'}^{X^nY^n}$. It follows from the result of Exercise 13.8.1 that $y^n \in T_{|\mathcal{X}|(\delta+\delta')}^{Y^n}$. □

13.10 Application: Shannon's Channel Capacity Theorem

We close the technical content of this chapter with a remarkable application of conditional typicality: Shannon's channel capacity theorem. As discussed in Section 2.2.3, this theorem is one of the central results of classical information theory, appearing in Shannon's seminal paper. The theorem establishes that the highest achievable rate for communication over many independent uses of a classical channel is equal to a simple function of the channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We begin by defining the information processing task and a corresponding (n, R, ϵ) channel code. It is helpful to recall Figure 2.4 depicting a general protocol for communication over a classical channel $\mathcal{N} \equiv p_{Y|X}(y|x)$. Before communication begins, the sender Alice and receiver Bob have already established a codebook $\{x^n(m)\}_{m\in\mathcal{M}}$, where each codeword $x^n(m)$ corresponds to a message m that Alice might wish to send to Bob. If Alice wishes to send message m, she inputs the codeword $x^n(m)$ to the IID channel $\mathcal{N}^n \equiv p_{Y^n|X^n}(y^n|x^n)$. More formally, her map is some encoding $E^n : \mathcal{M} \to \mathcal{X}^n$. She then exploits n uses of the channel to send $x^n(m)$. Bob receives some sequence y^n from the output of the channel, and he performs a decoding $D^n : \mathcal{Y}^n \to \mathcal{M}$ in order to recover the message m that Alice transmits. The rate R of the code is equal to $\log_2|\mathcal{M}|/n$, measured in bits per channel use. The probability of error p_e for an (n, R, ϵ) channel code is bounded from above as

$$p_e \equiv \max_m \Pr\{D^n(\mathcal{N}^n(E^n(m))) \neq m\} \leq \epsilon. \tag{13.173}$$

A communication rate R is achievable if there exists an $(n, R-\delta, \epsilon)$ channel code for all $\epsilon, \delta > 0$ and sufficiently large n. The channel capacity $C(\mathcal{N})$ of $\mathcal{N}$ is the supremum of all achievable rates. We can now state Shannon's channel capacity theorem:

Theorem 13.10.1 (Shannon Channel Capacity)**.** *The maximum mutual information $I(\mathcal{N})$ is equal to the capacity $C(\mathcal{N})$ of a channel $\mathcal{N} \equiv p_{Y|X}(y|x)$:*

$$C(\mathcal{N}) = I(\mathcal{N}) \equiv \max_{p_X(x)} I(X;Y). \tag{13.174}$$

Proof. The proof consists of two parts. The first part, known as the direct coding theorem, demonstrates that the RHS $\leq$ LHS. That is, there is a sequence of channel codes with rate $I(\mathcal{N})$ that are achievable. The second part, known as the converse part, demonstrates that the LHS $\leq$ RHS. That is, it demonstrates that the rate on the RHS is optimal, and it is impossible to have achievable rates exceeding it. Here, we prove the direct coding theorem and hold off on proving the converse part until we reach the HSW theorem in Chapter 19 because the converse theorem there suffices as the converse part for this classical theorem. We have already outlined the proof of the direct coding theorem in Section 2.2.4, and it might be helpful at this point to review this section. In particular, the proof breaks down into three parts: random coding to establish the encoding, the decoding algorithm for the receiver, and the error analysis. We now give all of the technical details of the proof because this chapter has established all the tools that we need. **Code Construction.** Before communication begins, Alice and Bob agree upon a code by the following random selection procedure. For every message $m \in \mathcal{M}$, generate a codeword $x^n(m)$ IID according to the product distribution $p_{X^n}(x^n)$, where $p_X(x)$ is the distribution that maximizes $I(\mathcal{N})$. **Encoding.** If Alice wishes to send message m, she inputs the codeword $x^n(m)$ to the channels. **Decoding Algorithm.** After receiving the sequence y^n from the channel outputs, Bob tests whether y^n is in the typical set $T_\delta^{Y^n}$ corresponding to the distribution $p_Y(y) \equiv \sum_x p_{Y|X}(y|x)p_X(x)$. If it is not, then he reports an error. He then tests if there is some message m such that the sequence y^n is in the conditionally typical set $T_\delta^{Y^n|x^n(m)}$. If m is the unique message such that $y^n \in$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$T_\delta^{Y^n|x^n(m)}$, then he declares m to be the transmitted message. If there is no message m such that $y^n \in T_\delta^{Y^n|x^n(m)}$ or multiple messages m' such that $y^n \in T_\delta^{Y^n|x^n(m')}$, then he reports an error. Observe that the decoder is a function of the channel, so that we might say that we construct channel codes "from the channel."**Error Analysis.** As discussed in the above decoding algorithm, there are three kinds of errors that can occur in this communication scheme when Alice sends the codeword $x^n(m)$ over the channels:

$\mathcal{E}_0(m)$**:** The event that the channel output y^n is not in the typical set $T_\delta^{Y^n}$.

$\mathcal{E}_1(m)$**:** The event that the channel output y^n is in $T_\delta^{Y^n}$ but not in the conditionally typical set $T_\delta^{Y^n|x^n(m)}$.

$\mathcal{E}_2(m)$**:** The event that the channel output y^n is in $T_\delta^{Y^n}$ but it is in the conditionally typical set for some other message:

$$\left\{y^n \in T_\delta^{Y^n}\right\} \text{ and } \left\{\exists m' : y^n \in T_\delta^{Y^n|x^n(m')}\right\}. \tag{13.175}$$

For each of the above events, we can exploit indicator functions in order to simplify the error analysis (we are also doing this to help build a bridge between this classical proof and the packing lemma approach for the quantum case in Chapter 15—projectors in some sense replace indicator functions later on). Recall that an indicator function $I_{\mathcal{A}}(x)$ is equal to one if $x \in \mathcal{A}$ and equal to zero otherwise. So the following three functions being equal to one or larger then corresponds to error events $\mathcal{E}_0(m)$, $\mathcal{E}_1(m)$, and $\mathcal{E}_2(m)$, respectively:

$$1 - I_{T_\delta^{Y^n}}(y^n), \tag{13.176}$$

$$I_{T_\delta^{Y^n}}(y^n)\left(1 - I_{T_\delta^{Y^n|x^n(m)}}(y^n)\right), \tag{13.177}$$

$$\sum_{m' \neq m} I_{T_\delta^{Y^n}}(y^n) I_{T_\delta^{Y^n|x^n(m')}}(y^n). \tag{13.178}$$

Recall from Section 2.2.4 that it is helpful to analyze the expectation of the average error probability, where the expectation is with respect to the random selection of the code and the average is with respect to a uniformly random choice of the message m. That is, we analyze

$$\mathbb{E}_{X^n}\left\{\frac{1}{|\mathcal{M}|}\sum_m \Pr\{\mathcal{E}_0(m) \cup \mathcal{E}_1(m) \cup \mathcal{E}_2(m)\}\right\}. \tag{13.179}$$

(The notation $\mathbb{E}_{X^n}$ implicitly indicates an expectation over all codewords.) Our first "move" is to exchange the expectation and the sum:

$$\frac{1}{|\mathcal{M}|}\sum_m \mathbb{E}_{X^n(m)}\{\Pr\{\mathcal{E}_0(m) \cup \mathcal{E}_1(m) \cup \mathcal{E}_2(m)\}\}. \tag{13.180}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Since all codewords are selected in the same way (randomly and independently of the message m), it suffices to analyze $\mathbb{E}_{X^n(m)}\{\Pr\{\mathcal{E}_0(m) \cup \mathcal{E}_1(m) \cup \mathcal{E}_2(m)\}\}$ for just a single message m. So we can then apply the union bound:

$$\begin{aligned}\mathbb{E}_{X^n(m)}\{\Pr\{\mathcal{E}_0(m) \cup \mathcal{E}_1(m) \cup \mathcal{E}_2(m)\}\} \\ \leq \mathbb{E}_{X^n(m)}\{\Pr\{\mathcal{E}_0(m)\}\} + \mathbb{E}_{X^n(m)}\{\Pr\{\mathcal{E}_1(m)\}\} + \mathbb{E}_{X^n(m)}\{\Pr\{\mathcal{E}_2(m)\}\}. \quad (13.181)\end{aligned}$$

We now analyze each error individually. By exploiting the indicator function from (13.176), we have that

$$\begin{aligned}&\mathbb{E}_{X^n(m)}\{\Pr\{\mathcal{E}_0(m)\}\} \\ &= \mathbb{E}_{X^n(m)}\Big\{\mathbb{E}_{Y^n|X^n(m)}\Big\{1 - I_{T_\delta^{Y^n}}(Y^n)\Big\}\Big\} && (13.182)\\ &= 1 - \mathbb{E}_{X^n(m),Y^n}\Big\{I_{T_\delta^{Y^n}}(Y^n)\Big\} && (13.183)\\ &= 1 - \mathbb{E}_{Y^n}\Big\{I_{T_\delta^{Y^n}}(Y^n)\Big\} && (13.184)\\ &= \Pr\big\{Y^n \notin T_\delta^{Y^n}\big\} && (13.185)\\ &\leq \epsilon, && (13.186)\end{aligned}$$

where in the last line we have exploited the high probability property of the typical set $T_\delta^{Y^n}$. In the above, we are also exploiting the fact that $\mathbb{E}\{I_{\mathcal{A}}\} = \Pr\{\mathcal{A}\}$. By exploiting the indicator function from (13.177), we have that

$$\begin{aligned}&\mathbb{E}_{X^n(m)}\{\Pr\{\mathcal{E}_1(m)\}\} \\ &= \mathbb{E}_{X^n(m)}\Big\{\mathbb{E}_{Y^n|X^n(m)}\Big\{I_{T_\delta^{Y^n}}(Y^n)\Big(1 - I_{T_\delta^{Y^n|X^n(m)}}(Y^n)\Big)\Big\}\Big\} && (13.187)\\ &\leq \mathbb{E}_{X^n(m)}\Big\{\mathbb{E}_{Y^n|X^n(m)}\Big\{1 - I_{T_\delta^{Y^n|X^n(m)}}(Y^n)\Big\}\Big\} && (13.188)\\ &= 1 - \mathbb{E}_{X^n(m)}\Big\{\mathbb{E}_{Y^n|X^n(m)}\Big\{I_{T_\delta^{Y^n|X^n(m)}}(Y^n)\Big\}\Big\} && (13.189)\\ &= \mathbb{E}_{X^n(m)}\Big\{\Pr_{Y^n|X^n(m)}\Big\{Y^n \notin T_\delta^{Y^n|X^n(m)}\Big\}\Big\} && (13.190)\\ &\leq \epsilon, && (13.191)\end{aligned}$$

where in the last line we have exploited the high probability property of the conditionally typical set $T_\delta^{Y^n|X^n(m)}$. We finally consider the probability of the last kind of error by

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

exploiting the indicator function in (13.178):

$$\begin{aligned}
&\mathbb{E}_{X^n(m)}\{\Pr\{\mathcal{E}_2(m)\}\} \\
&\leq \sum_{m' \neq m} \mathbb{E}_{X^n(m),X^n(m'),Y^n}\left\{ I_{T_\delta^{Y^n}}(y^n) I_{T_\delta^{Y^n|x^n(m')}}(y^n) \right\} && (13.192) \\
&= \sum_{m' \neq m} \sum_{x^n(m),x^n(m'),y^n} p_{X^n}(x^n(m)) p_{X^n}(x^n(m')) \times \\
&\qquad p_{Y^n|X^n}(y^n|x^n(m)) I_{T_\delta^{Y^n}}(y^n) I_{T_\delta^{Y^n|x^n(m')}}(y^n) && (13.193) \\
&= \sum_{m' \neq m} \sum_{x^n(m'),y^n} p_{X^n}(x^n(m')) p_{Y^n}(y^n) I_{T_\delta^{Y^n}}(y^n) I_{T_\delta^{Y^n|x^n(m')}}(y^n) && (13.194)
\end{aligned}$$

The first inequality is from the union bound, and the first equality follows from the way that we select the random code: for every message m, the codewords are selected independently and randomly according to p_{X^n} so that the distribution for the joint random variable $X^n(m)X^n(m')Y^n$ is

$$p_{X^n}(x^n(m)) p_{X^n}(x^n(m')) p_{Y^n|X^n}(y^n|x^n(m)). \tag{13.195}$$

The second equality follows from marginalizing over $X^n(m)$. Continuing, we have

$$\begin{aligned}
&\leq 2^{-n[H(Y)-\delta]} \sum_{m' \neq m} \sum_{x^n(m'),y^n} p_{X^n}(x^n(m')) I_{T_\delta^{Y^n|x^n(m')}}(y^n) && (13.196) \\
&= 2^{-n[H(Y)-\delta]} \sum_{m' \neq m} \sum_{x^n(m')} p_{X^n}(x^n(m')) \sum_{y^n} I_{T_\delta^{Y^n|x^n(m')}}(y^n) && (13.197) \\
&\leq 2^{-n[H(Y)-\delta]} 2^{n[H(Y|X)+\delta]} \sum_{m' \neq m} \sum_{x^n(m')} p_{X^n}(x^n(m')) && (13.198) \\
&\leq |\mathcal{M}| 2^{-n[I(X;Y)-2\delta]}. && (13.199)
\end{aligned}$$

The first inequality follows from the bound $p_{Y^n}(y^n) I_{T_\delta^{Y^n}}(y^n) \leq 2^{-n[H(Y)-\delta]}$ that holds for typical sequences. The second inequality follows from the cardinality bound $\left|T_\delta^{Y^n|x^n(m')}\right| \leq 2^{n[H(Y|X)+\delta]}$ on the conditionally typical set. The last inequality follows because

$$\sum_{x^n(m')} p_{X^n}(x^n(m')) = 1, \tag{13.200}$$

$|\mathcal{M}|$ is an upper bound on $\sum_{m' \neq m} 1$, and by the identity $I(X;Y) = H(Y) - H(Y|X)$. Thus, we can make this error arbitrarily small by choosing the message set size $|\mathcal{M}| = 2^{n[I(X;Y)-3\delta]}$. Putting everything together, we have the following bound on (13.179)

$$\epsilon' \equiv 2\epsilon + 2^{-n\delta}, \tag{13.201}$$

as long as we choose the message set size as given above. It follows that there exists a particular code with the same error bound on its average error probability. We can then

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

exploit an expurgation argument as discussed in Section 2.2.4 to convert an average error bound into a maximal one. Thus, we have shown the achievability of an $(n, C(\mathcal{N}) - \delta', \epsilon')$ channel code for all $\delta', \epsilon' > 0$ and sufficiently large n. Finally, as a simple observation, our proof above does not rely on whether the definition of conditional typicality employed is weak or strong. □

13.11 Concluding Remarks

This chapter deals with many different definitions and flavors of typicality in the classical world, but the essential theme is Shannon's central insight—the application of the law of large numbers in information theory. Our main goal in information theory is to analyze the probability of error in the transmission or compression of information. Thus, we deal with probabilities and we do not care much what happens for all sequences, but we instead only care what happens for the likely sequences. This frame of mind immediately leads to the definition of a typical sequence and to a simple scheme for the compression of information—keep only the typical sequences and performance is optimal in the asymptotic limit. Despite the seemingly different nature of quantum information when compared to its classical counterpart, the intuition developed in this chapter carries over to the quantum world in the next chapter where we define several different notions of quantum typicality.

13.12 History and Further Reading

The book of Cover and Thomas contains a great presentation of typicality in the classical case [57]. The proof of Property 13.7.5 is directly from the Cover and Thomas book. Berger introduced strong typicality [37], and Csiszár and Körner systematically developed it [58]. Other useful books on information theory are that of Berger [36] and Yeung [273]. There are other notions of typicality which are useful, including those presented in Ref. [89] and Ref. [264]. Our proof of Shannon's channel capacity theorem is similar to that in Savov's thesis [211].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 14

Quantum Typicality

This chapter marks the beginning of our study of the asymptotic theory of quantum information, where we develop the technical tools underpinning this theory. The intuition for it is similar to the intuition we developed in the previous chapter on typical sequences, but we will find some important differences between the classical and quantum cases.

So far, there is not a single known information processing task in quantum Shannon theory where the tools from this chapter are not helpful in proving the achievability part of a coding theorem. For the most part, we can straightforwardly import many of the ideas from the previous chapter about typical sequences for use in the asymptotic theory of quantum information. Though, one might initially think that there are some obstacles to doing so. For example, what is the analogy of a quantum information source? Once we have established this notion, how would we determine if a state emitted from a quantum information source is a typical state? In the classical case, a simple way of determining typicality is to inspect all of the bits in the sequence. But there is a problem with this approach in the quantum domain—"looking at quantum bits" is equivalent to performing a measurement and doing so destroys delicate superpositions that we would want to preserve in any subsequent quantum information processing task.

So how can we get around the aforementioned problem and construct a useful notion of quantum typicality? Well, we should not be so destructive in determining the answer to a question when it has only two possible answers. After all, we are only asking "Is the state typical or not?", and we can be a bit more delicate in the way that we ask this question. As an analogy, suppose Bob is curious to determine whether Alice could join him for dinner at a nice restaurant on the beach. He would likely just phone her and politely ask, "Sweet Alice, are you available for a lovely oceanside dinner?", as opposed to barging into her apartment, probing through all of her belongings in search of her calendar, and demanding that she join him if she is available. This latter infraction would likely disturb her so much that she would never speak to him again (and what would become of quantum Shannon theory without these two communicating!). It is the same with quantum information—we must be gentle when handling quantum states. Otherwise, we will disturb the state so much that it will not be useful in any future quantum information processing task.

We can gently ask a binary question to a quantum system by constructing an incomplete measurement with only two outcomes. If one outcome has a high probability of occurring, then we do not learn much about the state after learning this outcome, and thus we would expect that this inquiry does not disturb the state very much. For the case above, we can formulate the question, "Is the state typical or not?" as a binary measurement that returns only the answer to this question and no more information. Since it is highly likely that the state is indeed a typical state, we would expect this inquiry not to disturb the state very much, and we could use it for further quantum information processing tasks. This is the essential content of this chapter, and there are several technicalities necessary to provide a rigorous underpinning.

We structure this chapter as follows. We first discuss the notion of a typical subspace (the quantum analogy of the typical set). We can employ weak or strong notions of typicality in the definition of quantum typicality. Section 14.2 then discusses conditional quantum typicality, a form of quantum typicality that applies to quantum states chosen randomly according to a classical sequence. We end this chapter with a brief discussion of the method of types for quantum systems. All of these developments are important for understanding the asymptotic nature of quantum information and for determining the ultimate limits of storage and transmission with quantum media.

14.1 The Typical Subspace

Our first task is to establish the notion of a quantum information source. It is analogous to the notion of a classical information source, in the sense that the source randomly outputs a quantum state according to some probability distribution, but the states that it outputs do not necessarily have to be distinguishable as in the classical case.

Definition 14.1.1 (Quantum Information Source)**.** *A quantum information source is some device that randomly emits pure qudit states in a Hilbert space $\mathcal{H}_X$ of size $|\mathcal{X}|$.*

We use the symbol X to denote the quantum system for the quantum information source in addition to denoting the Hilbert space in which the state lives. Suppose that the source outputs states $|\psi_y\rangle$ randomly according to some probability distribution $p_Y(y)$. Note that the states $|\psi_y\rangle$ do not necessarily have to form an orthonormal set. Then the density operator ρ^X of the source is the expected state emitted:

$$\rho^X \equiv \mathbb{E}_Y\{|\psi_Y\rangle\langle\psi_Y|\} = \sum_y p_Y(y)|\psi_y\rangle\langle\psi_y|. \tag{14.1}$$

There are many decompositions of a density operator as a convex sum of rank-one projectors (and the above decomposition is one such example), but perhaps the most important decomposition is a spectral decomposition of the density operator ρ:

$$\rho^X = \sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x|^X. \tag{14.2}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above states $|x\rangle^X$ are eigenvectors of ρ^X and form a complete orthonormal basis for Hilbert space $\mathcal{H}_X$, and the non-negative, convex real numbers $p_X(x)$ are the eigenvalues of ρ^X.

We have written the states $|x\rangle^X$ and the eigenvalues $p_X(x)$ in a suggestive notation because it is actually possible to think of our quantum source as a classical information source—the emitted states $\{|x\rangle^X\}_{x\in\mathcal{X}}$ are orthonormal and each corresponding eigenvalue $p_X(x)$ acts as a probability for choosing $|x\rangle^X$. We can say that our source is classical because it is emitting the orthogonal, and thus distinguishable, states $|x\rangle^X$ with probability $p_X(x)$. This description is equivalent to the ensemble $\{p_Y(y), |\psi_y\rangle\}_y$ because the two ensembles lead to the same density operator (recall that two ensembles that have the same density operator are essentially equivalent because they lead to the same probabilities for outcomes of any measurement performed on the system). Our quantum information source then corresponds to the pure-state ensemble:

$$\left\{p_X(x), |x\rangle^X\right\}_{x\in\mathcal{X}}. \tag{14.3}$$

Recall that the von Neumann entropy $H(X)$ of the density operator ρ^X is as follows (Definition 11.1.1):

$$H(X)_\rho \equiv -\text{Tr}\{\rho^X \log \rho^X\}. \tag{14.4}$$

It is straightforward to show that the von Neumann entropy $H(X)_\rho$ is equal to the Shannon entropy $H(X)$ of a random variable X with distribution $p_X(x)$ because the basis states $|x\rangle^X$ are orthonormal.

Suppose now that the quantum information source emits a large number n of random quantum states so that the density operator describing the emitted state is as follows:

$$\rho^{X^n} \equiv \underbrace{\rho^{X_1}\otimes\cdots\otimes\rho^{X_n}}_{n \text{ times}} = \left(\rho^X\right)^{\otimes n}. \tag{14.5}$$

The labels $X_1, \ldots, X_n$ denote the Hilbert spaces in which the different quantum systems live, but the density operator is the same for each Hilbert space $X_1, \ldots, X_n$ and is equal to ρ^X. The above description of a quantum source is within the independent and identically distributed (IID) setting for the quantum domain. The spectral decomposition of the state in (14.5) is as follows:

$$\rho^{X^n} = \sum_{x_1\in\mathcal{X}} p_X(x_1)|x_1\rangle\langle x_1|^{X_1}\otimes\cdots\otimes\sum_{x_n\in\mathcal{X}} p_X(x_n)|x_n\rangle\langle x_n|^{X_n} \tag{14.6}$$

$$= \sum_{x_1,\cdots,x_n\in\mathcal{X}} p_X(x_1)\cdots p_X(x_n)|x_1\rangle\cdots|x_n\rangle\langle x_1|\cdots\langle x_n|^{X_1,\ldots,X_n} \tag{14.7}$$

$$= \sum_{x^n\in\mathcal{X}^n} p_{X^n}(x^n)|x^n\rangle\langle x^n|^{X^n}, \tag{14.8}$$

where we employ the shorthand:

$$p_{X^n}(x^n) \equiv p_X(x_1)\cdots p_X(x_n), \qquad |x^n\rangle^{X^n} \equiv |x_1\rangle^{X_1}\cdots|x_n\rangle^{X_n}. \tag{14.9}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above quantum description of the density operator is essentially equivalent to the classical picture of n realizations of random variable X with each eigenvalue $p_{X_1}(x_1)\cdots p_{X_n}(x_n)$ acting as a probability because the set of states $\{|x_1\rangle\cdots|x_n\rangle^{X_1,\ldots,X_n}\}_{x_1,\cdots,x_n\in\mathcal{X}}$ is an orthonormal set.

We can now "quantize" or extend the notion of typicality to the quantum information source. The definitions follow directly from the classical definitions in Chapter 13. The quantum definition of typicality can employ either the weak notion as in Definition 13.2.3 or the strong notion as in Definition 13.7.2. We do not distinguish the notation for a typical subspace and a typical set because it should be clear from context which kind of typicality we are employing.

Definition 14.1.2 (Typical Subspace)**.** *The δ-typical subspace $T_\delta^{X^n}$ is a subspace of the full Hilbert space $X_1, \ldots, X_n$ and is associated with many copies of a density operator, such as the one in (14.2). It is spanned by states $|x^n\rangle^{X^n}$ whose corresponding classical sequences x^n are δ-typical:*

$$T_\delta^{X^n} \equiv \text{span}\Big\{|x^n\rangle^{X^n} : x^n \in T_\delta^{X^n}\Big\}, \tag{14.10}$$

where it is implicit that the typical subspace $T_\delta^{X^n}$ on the LHS is with respect to a density operator ρ and the typical set $T_\delta^{X^n}$ on the RHS is with respect to the distribution $p_X(x)$ from the spectral decomposition of ρ in (14.2). We could also denote the typical subspace as $T_{\rho,\delta}^{X^n}$ if we would like to make the dependence of the space on ρ more explicit.

14.1.1 The Typical Subspace Measurement

The definition of the typical subspace (Definition 14.1.2) gives a way to divide up the Hilbert space of n qudits into two subspaces: the typical subspace and the atypical subspace. The properties of the typical subspace are similar to what we found for the properties of typical sequences. That is, the typical subspace is exponentially smaller than the full Hilbert space of n qudits, yet it contains nearly all of the probability (in a sense that we show below). The intuition for these properties of the typical subspace is the same as it is classically, as depicted in Figure 13.2, once we have a spectral decomposition of a density operator.

The *typical projector* is a projector onto the typical subspace, and the complementary projector projects onto the atypical subspace. These projectors play an important operational role in quantum Shannon theory because we can construct a quantum measurement from them. That is, this measurement is the best way of asking the question, "Is the state typical or not?" because it minimally disturbs the state while still retrieving this one bit of information.

Definition 14.1.3 (Typical Projector)**.** *Let $\Pi_\delta^{X^n}$ denote the typical projector for the typical subspace of a density operator ρ^X with spectral decomposition in (14.2). It is a projector onto the typical subspace:*

$$\Pi_\delta^{X^n} \equiv \sum_{x^n\in T_\delta^{X^n}} |x^n\rangle\langle x^n|^{X^n}, \tag{14.11}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where it is implicit that the x^n below the summation is a classical sequence in the typical set $T_\delta^{X^n}$, and the state $|x^n\rangle$ is a quantum state given in (14.9) and associated with the the classical sequence x^n via the spectral decomposition of ρ in (14.2). We can also denote the typical projector as $\Pi_{\rho,\delta}^{X^n}$ if we would like to make its dependence on ρ explicit.

The action of multiplying the density operator ρ^{X^n} by the typical projector $\Pi_\delta^{X^n}$ is to select out all the basis states of ρ^{X^n} that are in the typical subspace and form a "sliced" operator $\tilde{\rho}^{X^n}$ that is close to the original density operator ρ^{X^n}:

$$\tilde{\rho}^{X^n} \equiv \Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n} = \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n) |x^n\rangle\langle x^n|^{X^n}. \tag{14.12}$$

That is, the effect of projecting a state onto the typical subspace $T_\delta^{X^n}$ is to "slice" out any component of the state ρ^{X^n} that does not lie in the typical subspace $T_\delta^{X^n}$.

Exercise 14.1.1 Show that the typical projector $\Pi_\delta^{X^n}$ commutes with the density operator ρ^{X^n}:

$$\rho^{X^n} \Pi_\delta^{X^n} = \Pi_\delta^{X^n} \rho^{X^n}. \tag{14.13}$$

The typical projector allows us to formulate an operational method for delicately asking the question: "Is the state typical or not?" We can construct a quantum measurement that consists of two outcomes: the outcome "1" reveals that the state is in the typical subspace, and "0" reveals that it is not. This typical subspace measurement is often one of the first important steps in most protocols in quantum Shannon theory.

Definition 14.1.4 (Typical Subspace Measurement)**.** *The following map is a quantum instrument (see Section 4.4.7) that realizes the typical subspace measurement:*

$$\sigma \to \left(I - \Pi_\delta^{X^n}\right)\sigma\left(I - \Pi_\delta^{X^n}\right) \otimes |0\rangle\langle 0| + \Pi_\delta^{X^n} \sigma \Pi_\delta^{X^n} \otimes |1\rangle\langle 1|, \tag{14.14}$$

where σ is some quantum state living in the Hilbert space X^n. It associates a classical register with the outcome of the measurement—the value of the classical register is $|0\rangle$ for the support of the state σ that is not in the typical subspace, and it is equal to $|1\rangle$ for the support of the state σ that is in the typical subspace.

The implementation of a typical subspace measurement is currently far from the reality of what is experimentally accessible if we would like to have the measure concentration effects necessary for proving many of the results in quantum Shannon theory. Recall from Figure 13.1 that we required a sequence of about a million bits in order to have the needed measure concentration effects. We would need a similar number of qubits emitted from a quantum information source, and furthermore, we would require the ability to perform noiseless coherent operations over about a million or more qubits in order to implement the typical subspace measurement. Such a daunting requirement firmly places quantum Shannon theory as a "highly theoretical theory," rather than being a theory that can make close connection to current experimental practice.[1]

[1] We should note that this was certainly the case as well for information theory when Claude Shannon developed it in 1948, but in the many years since then, there has been much progress in the development of practical classical codes for achieving the classical capacity of a classical channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

14.1.2 The Difference between the Typical Set and the Typical Subspace

We now offer a simple example to discuss the difference between the classical viewpoint associated with the typical set and the quantum viewpoint associated with the typical subspace. Suppose that a quantum information source emits the state $|+\rangle$ with probability 1/2 and it emits the state $|0\rangle$ with probability 1/2. For the moment, let us ignore the fact that the two states $|+\rangle$ and $|0\rangle$ are not perfectly distinguishable and instead suppose that they are. Then it would turn out that nearly every sequence emitted from this source is a typical sequence because the distribution of the source is uniform. Recall that the typical set has size roughly equal to $2^{nH(X)}$, and in this case, the entropy of the distribution $\left(\frac{1}{2}, \frac{1}{2}\right)$ is equal to one bit. Thus the size of the typical set is roughly the same as the size of the set of all sequences for this distribution because $2^{nH(X)} = 2^n$.

Now let us take into account the fact that the states $|+\rangle$ and $|0\rangle$ are not perfectly distinguishable and use the prescription given in Definition 14.1.2 for the typical subspace. The density operator of the above ensemble is as follows:

$$\frac{1}{2}|+\rangle\langle+| + \frac{1}{2}|0\rangle\langle 0| = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix}, \tag{14.15}$$

where its matrix representation is with respect to the computational basis. The spectral decomposition of the density operator is

$$\cos^2(\pi/8)|\psi_0\rangle\langle\psi_0| + \sin^2(\pi/8)|\psi_1\rangle\langle\psi_1|, \tag{14.16}$$

where the states $|\psi_0\rangle$ and $|\psi_1\rangle$ are orthogonal, and thus distinguishable from one another. The quantum information source that outputs $|0\rangle$ and $|+\rangle$ with equal probability is thus equivalent to a source that outputs $|\psi_0\rangle$ with probability $\cos^2(\pi/8)$ and $|\psi_1\rangle$ with probability $\sin^2(\pi/8)$.

We construct the projector onto the typical subspace by taking sums of typical strings of the states $|\psi_0\rangle\langle\psi_0|$ and $|\psi_1\rangle\langle\psi_1|$ rather than the states $|0\rangle\langle 0|$ and $|+\rangle\langle+|$, where typicality is with respect to the distribution $\left(\cos^2(\pi/8), \sin^2(\pi/8)\right)$. The dimension of the typical subspace corresponding to the quantum information source is far different from the size of the aforementioned typical set corresponding to the distribution $(1/2, 1/2)$. It is roughly equal to $2^{0.6n}$ because the entropy of the distribution $\left(\cos^2(\pi/8), \sin^2(\pi/8)\right)$ is about 0.6 bits. This stark contrast in the sizes has to do with the non-orthogonality of the states from the original description of the ensemble. That is, non-orthogonality of states in an ensemble implies that the size of the typical subspace can potentially be dramatically smaller than the size of the typical set corresponding to the distribution of the states in the ensemble. This result of course has implications for the compressibility of quantum information, and we will discuss these ideas in more detail in Chapter 17. For now, we continue with the technical details of typical subspaces.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

14.1.3 Properties of the Typical Subspace

The typical subspace $T_\delta^{X^n}$ enjoys several useful properties that are "quantized" versions of the typical sequence properties:

Property 14.1.1 (Unit Probability) Suppose that we perform a typical subspace measurement of a state ρ^{X^n}. Then the probability that the quantum state ρ^{X^n} is in the typical subspace $T_\delta^{X^n}$ approaches one as n becomes large:

$$\forall \epsilon > 0 \qquad \text{Tr}\{\Pi_\delta^{X^n} \rho^{X^n}\} \geq 1 - \epsilon \quad \text{for sufficiently large } n, \tag{14.17}$$

where $\Pi_\delta^{X^n}$ is the typical subspace projector from Definition 14.1.3.

Property 14.1.2 (Exponentially Small Dimension) The dimension $\dim(T_\delta^{X^n})$ of the δ-typical subspace is exponentially smaller than the dimension $|\mathcal{X}|^n$ of the entire space of quantum states when the output of the quantum information source is not maximally mixed. We formally state this property as follows:

$$\text{Tr}\{\Pi_\delta^{X^n}\} \leq 2^{n(H(X)+c\delta)}, \tag{14.18}$$

where c is some constant that depends on whether we employ the weak or strong notion of typicality. We can also lower bound the dimension $\dim(T_\delta^{X^n})$ of the δ-typical subspace when n is sufficiently large:

$$\forall \epsilon > 0 \qquad \text{Tr}\{\Pi_\delta^{X^n}\} \geq (1-\epsilon) 2^{n(H(X)-c\delta)} \quad \text{for sufficiently large } n. \tag{14.19}$$

Property 14.1.3 (Equipartition) The operator $\Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n}$ corresponds to a "slicing" of the density operator ρ^{X^n} where we slice out and keep only the part with support in the typical subspace. We can then bound all of the eigenvalues of the sliced operator $\Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n}$ as follows:

$$2^{-n(H(X)+c\delta)} \Pi_\delta^{X^n} \leq \Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n} \leq 2^{-n(H(X)-c\delta)} \Pi_\delta^{X^n}. \tag{14.20}$$

The above inequality is an operator inequality. It is a statement about the eigenvalues of the operators $\Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n}$ and $\Pi_\delta^{X^n}$, and these operators have the same eigenvectors because they commute. Therefore, the above inequality is equivalent to the following inequality that applies in the classical case:

$$\forall x^n \in T_\delta^{X^n} : 2^{-n(H(X)+c\delta)} \leq p_{X^n}(x^n) \leq 2^{-n(H(X)-c\delta)}. \tag{14.21}$$

This equivalence holds because each probability $p_{X^n}(x^n)$ is an eigenvalue of $\Pi_\delta^{X^n} \rho^{X^n} \Pi_\delta^{X^n}$.

The dimension $\dim(T_\delta^{X^n})$ of the δ-typical subspace is approximately equal to the dimension $|\mathcal{X}|^n$ of the entire space only when the density operator of the quantum information source is maximally mixed because

$$\text{Tr}\{\Pi_\delta^{X^n}\} \leq |\mathcal{X}|^n \cdot 2^{n\delta} \simeq |\mathcal{X}|^n. \tag{14.22}$$

The proofs of the above properties are essentially identical to those from the classical case in Sections 13.7.3 and 13.9.3, regardless of whether we employ a weak or strong notion of quantum typicality. We leave the proofs as the three exercises below.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 14.1.2 Prove the Unit Probability Property of the δ-typical subspace (Property 14.1.1). First show that the probability that many copies of a density operator is in the δ-typical subspace is equal to the probability that a random sequence is δ-typical:

$$\text{Tr}\{\Pi_\delta^{X^n}\rho^{X^n}\} = \Pr\{X^n \in T_\delta^{X^n}\}. \tag{14.23}$$

Exercise 14.1.3 Prove the Exponentially Small Dimension Property of the δ-typical subspace (Property 14.1.2). First show that the trace of the typical projector $\Pi_\delta^{X^n}$ is equal to the dimension of the typical subspace $T_\delta^{X^n}$:

$$\dim\left(T_\delta^{X^n}\right) = \text{Tr}\{\Pi_\delta^{X^n}\}. \tag{14.24}$$

Then prove the property.

Exercise 14.1.4 Prove the Equipartition Property of the δ-typical subspace (Property 14.1.3). First show that

$$\Pi_\delta^{X^n}\rho^{X^n}\Pi_\delta^{X^n} = \sum_{x^n \in T_\delta^{X^n}} p_{X^n}(x^n)|x^n\rangle\langle x^n|^{X^n}, \tag{14.25}$$

and then argue the proof.

The result of the following exercise shows that the sliced operator $\tilde{\rho}^{X^n} \equiv \Pi_\delta^{X^n}\rho^{X^n}\Pi_\delta^{X^n}$ is a good approximation to the original state ρ^{X^n} in the limit of many copies of the states, and it effectively gives a scheme for quantum data compression (more on this in Chapter 17).

Exercise 14.1.5 Use the Gentle Operator Lemma (Lemma 9.4.2) to show that ρ^{X^n} is $2\sqrt{\epsilon}$-close to the sliced operator $\tilde{\rho}^{X^n}$ when n is large:

$$\left\|\rho^{X^n} - \tilde{\rho}^{X^n}\right\|_1 \leq 2\sqrt{\epsilon}. \tag{14.26}$$

Use the Gentle Measurement Lemma (Lemma 9.4.1) to show that the sliced state

$$\left[\text{Tr}\{\Pi_\delta^{X^n}\rho^{X^n}\}\right]^{-1}\tilde{\rho}^{X^n} \tag{14.27}$$

is $2\sqrt{\epsilon}$-close in trace distance to $\tilde{\rho}^{X^n}$.

Exercise 14.1.6 Show that the purity $\text{Tr}\left\{\left(\tilde{\rho}^{X^n}\right)^2\right\}$ of the sliced state $\tilde{\rho}^{X^n}$ satisfies the following bound for sufficiently large n and any $\epsilon > 0$ (use weak quantum typicality):

$$(1-\epsilon)2^{-n(H(X)+\delta)} \leq \text{Tr}\left\{\left(\tilde{\rho}^{X^n}\right)^2\right\} \leq 2^{-n(H(X)-\delta)}. \tag{14.28}$$

Exercise 14.1.7 Show that the following bounds hold for the zero-norm and the ∞-norm of the sliced state $\tilde{\rho}^{X^n}$ for any $\epsilon > 0$ and sufficiently large n:

$$(1-\epsilon)2^{n(H(X)-\delta)} \leq \left\|\tilde{\rho}^{X^n}\right\|_0 \leq 2^{n(H(X)+\delta)}, \tag{14.29}$$

$$2^{-n(H(X)+\delta)} \leq \left\|\tilde{\rho}^{X^n}\right\|_\infty \leq 2^{-n(H(X)-\delta)}. \tag{14.30}$$

(Recall that the zero-norm of an operator is equal to the size of its support and that the infinity norm is equal to its maximum eigenvalue. Again use weak quantum typicality.)

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

14.1.4 The Typical Subspace for Bipartite or Multipartite States

Recall from Section 13.5 that two classical sequences x^n and y^n are weak jointly typical if the joint sample entropy of $x^n y^n$ is close to the joint entropy $H(X,Y)$ and if the sample entropies of the individual sequences are close to their respective marginal entropies $H(X)$ and $H(Y)$ (where the entropies are with respect to some joint distribution $p_{X,Y}(x,y)$). How would we then actually check that these conditions hold? The most obvious way is simply to look at the sequence $x^n y^n$, compute its joint sample entropy, compare this quantity to the true joint entropy, determine if the difference is under the threshold δ, and do the same for the marginal sequences. These two operations both commute in the sense that we can determine first if the marginals are typical and then if the joint sequence is typical or vice versa without any difference in which one we do first.

But such a commutation does not necessarily hold in the quantum world. The way that we determine whether a quantum state is typical is by performing a typical subspace measurement. If we perform a typical subspace measurement of the whole system followed by such a measurement on the marginals, the resulting state is not necessarily the same as if we performed the marginal measurements followed by the joint measurements. For this reason, the notion of weak joint typicality as given in Definition 13.5.3 does not really exist in general for the quantum case. Nevertheless, we still briefly overview how one would handle such a case and later give an example of a restricted class of states for which weak joint typicality holds.

Suppose that we have a quantum system in the mixed state ρ^{XY} shared between two parties X and Y. We can decompose the mixed state with the spectral theorem:

$$\rho^{XY} = \sum_{z \in \mathcal{Z}} p_Z(z) |\psi_z\rangle\langle\psi_z|^{XY}, \tag{14.31}$$

where the states $\{|\psi_z\rangle^{XY}\}_{z \in \mathcal{Z}}$ form an orthonormal basis for the joint quantum system XY and each of the states $|\psi_z\rangle^{XY}$ can be entangled in general.

We can consider the n^{th} extension $\rho^{X^n Y^n}$ of the above state and abbreviate its spectral decomposition as follows:

$$\rho^{X^n Y^n} \equiv \left(\rho^{XY}\right)^{\otimes n} = \sum_{z^n \in \mathcal{Z}^n} p_{Z^n}(z^n) |\psi_{z^n}\rangle\langle\psi_{z^n}|^{X^n Y^n}, \tag{14.32}$$

where

$$p_{Z^n}(z^n) \equiv p_Z(z_1) \cdots p_Z(z_n), \tag{14.33}$$

$$|\psi_{z^n}\rangle^{X^n Y^n} \equiv |\psi_{z_1}\rangle^{X_1 Y_1} \cdots |\psi_{z_n}\rangle^{X_n Y_n}. \tag{14.34}$$

This development immediately leads to the definition of the typical subspace for a bipartite state.

Definition 14.1.5 (Typical Subspace of a Bipartite State)**.** *The δ-typical subspace $T_\delta^{X^n Y^n}$ of ρ^{XY} is the space spanned by states $|\psi_{z^n}\rangle^{X^n Y^n}$ whose corresponding classical sequence z^n is*

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

in the typical set $T_\delta^{Z^n}$*:*

$$T_\delta^{X^nY^n} \equiv \text{span}\Big\{ |\psi_{z^n}\rangle^{X^nY^n} : z^n \in T_\delta^{Z^n} \Big\}. \tag{14.35}$$

The states $|\psi_{z^n}\rangle^{X^nY^n}$ *are from the spectral decomposition of* ρ^{XY}*, and the distribution to consider for typicality of the classical sequence* z^n *is* $p_Z(z)$ *from the spectral decomposition.*

Definition 14.1.6 (Typical Projector of a Bipartite State)**.** *Let* $\Pi_\delta^{X^nY^n}$ *denote the projector onto the typical subspace of* ρ^{XY}*:*

$$\Pi_\delta^{X^nY^n} \equiv \sum_{z^n \in T_\delta^{Z^n}} |\psi_{z^n}\rangle\langle\psi_{z^n}|^{X^nY^n}. \tag{14.36}$$

Thus, there is ultimately no difference between the typical subspace for a bipartite state and the typical subspace for a single-party state because the spectral decomposition gives a way for determining the typical subspace and the typical projector in both cases. Perhaps the only difference is a cosmetic one because XY denotes the bipartite system while Z indicates a random variable with a distribution given from the spectral decomposition. Finally, Properties 14.1.1-14.1.3 hold for quantum typicality of a bipartite state.

14.1.5 The Jointly Typical Subspace for Classical States

The notion of weak joint typicality may not hold in the general case, but it does hold for a special class of states that are completely classical. Suppose now that the mixed state ρ^{XY} shared between two parties X and Y has the following special form:

$$\rho^{XY} = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)(|x\rangle \otimes |y\rangle)(\langle x| \otimes \langle y|)^{XY} \tag{14.37}$$

$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)|x\rangle\langle x|^X \otimes |y\rangle\langle y|^Y, \tag{14.38}$$

where the states $\{|x\rangle^X\}_{x\in\mathcal{X}}$ and $\{|y\rangle^Y\}_{y\in\mathcal{Y}}$ form an orthonormal basis for the respective systems $\mathcal{X}$ and $\mathcal{Y}$. This state has only classical correlations because Alice and Bob can prepare it simply by local operations and classical communication. That is, Alice can sample from the distribution $p_{X,Y}(x,y)$ in her laboratory and send Bob the variable y. Furthermore, the states on X and Y locally form a distinguishable set.

We can consider the n^{th} extension $\rho^{X^nY^n}$ of the above state:

$$\rho^{X^nY^n} \equiv \left(\rho^{XY}\right)^{\otimes n} \tag{14.39}$$

$$= \sum_{x^n\in\mathcal{X}^n, y^n\in\mathcal{Y}^n} p_{X^n,Y^n}(x^n,y^n)(|x^n\rangle \otimes |y^n\rangle)(\langle x^n| \otimes \langle y^n|)^{X^nY^n} \tag{14.40}$$

$$= \sum_{x^n\in\mathcal{X}^n, y^n\in\mathcal{Y}^n} p_{X^n,Y^n}(x^n,y^n)|x^n\rangle\langle x^n|^{X^n} \otimes |y^n\rangle\langle y^n|^{Y^n}. \tag{14.41}$$

This development immediately leads to the definition of the weak jointly typical subspace for this special case.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Definition 14.1.7 (Jointly Typical Subspace)**.** *The weak δ-jointly-typical subspace $T_\delta^{X^nY^n}$ is the space spanned by states $|x^n\rangle|y^n\rangle^{X^nY^n}$ whose corresponding classical sequence x^ny^n is in the jointly-typical set:*

$$T_\delta^{X^nY^n} \equiv \text{span}\Big\{|x^n\rangle|y^n\rangle^{X^nY^n} : x^ny^n \in T_\delta^{X^nY^n}\Big\}. \tag{14.42}$$

Definition 14.1.8 (Jointly Typical Projector)**.** *Let $\Pi_\delta^{X^nY^n}$ denote the jointly-typical projector. It is the projector onto the jointly-typical subspace:*

$$\Pi_\delta^{X^nY^n} \equiv \sum_{x^n,y^n \in T_\delta^{X^nY^n}} |x^n\rangle\langle x^n|^{X^n} \otimes |y^n\rangle\langle y^n|^{Y^n}. \tag{14.43}$$

Properties of the Jointly Typical Projector

Properties 14.1.1-14.1.3 apply to the jointly-typical subspace $T_\delta^{X^nY^n}$ because it is a typical subspace. The following property, analogous to Property 13.5.4 for classical joint typicality, holds because the state ρ^{XY} has the special form in (14.37):

Property 14.1.4 (Probability of Joint Typicality) Consider the following marginal density operators:

$$\rho^{X^n} \equiv \text{Tr}_{Y^n}\{\rho^{X^nY^n}\}, \qquad \rho^{Y^n} \equiv \text{Tr}_{X^n}\{\rho^{X^nY^n}\}. \tag{14.44}$$

Let us define $\rho^{\tilde{X}^n\tilde{Y}^n}$ as the following density operator:

$$\rho^{\tilde{X}^n\tilde{Y}^n} \equiv \rho^{X^n} \otimes \rho^{Y^n} \neq \rho^{X^nY^n}. \tag{14.45}$$

The marginal density operators of $\rho^{\tilde{X}^n\tilde{Y}^n}$ are therefore equivalent to the marginal density operators of $\rho^{X^nY^n}$. Then we can bound the probability that the state $\rho^{\tilde{X}^n\tilde{Y}^n}$ lies in the typical subspace $T_\delta^{X^nY^n}$:

$$\text{Tr}\Big\{\Pi_\delta^{X^nY^n}\rho^{\tilde{X}^n\tilde{Y}^n}\Big\} \leq 2^{-n(I(X;Y)-3\delta)}. \tag{14.46}$$

Exercise 14.1.8 Prove the bound in Property 14.1.4:

$$\text{Tr}\Big\{\Pi_\delta^{X^nY^n}\rho^{\tilde{X}^n\tilde{Y}^n}\Big\} \leq 2^{-n(I(X;Y)-3\delta)}. \tag{14.47}$$

14.2 Conditional Quantum Typicality

The notion of conditional quantum typicality is somewhat similar to the notion of conditional typicality in the classical domain, but we again quickly notice some departures because different quantum states do not have to be perfectly distinguishable. The technical tools for conditional quantum typicality developed in this section are important for determining how

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

much public or private classical information we can send over a quantum channel (topics discussed in Chapters 19 and 22).

We first develop the notion of a conditional quantum information source. Consider a random variable X with probability distribution $p_X(x)$. Let $\mathcal{X}$ be the alphabet of the random variable, and let $|\mathcal{X}|$ denote its cardinality. We also associate a quantum system X with the random variable X and use an orthonormal set $\{|x\rangle\}_{x\in\mathcal{X}}$ to represent its realizations. We again label the elements of the alphabet $\mathcal{X}$ as $\{x\}_{x\in\mathcal{X}}$.

Suppose we generate a realization x of random variable X according to its distribution $p_X(x)$, and we follow by generating a random quantum state according to some conditional distribution. This procedure then gives us a set of $|\mathcal{X}|$ quantum information sources (each of them are as in Definition 14.1.1). We index them by the classical index x, and the quantum information source has expected density operator ρ_x^B if the emitted classical index is x. Furthermore, we impose the constraint that each ρ_x^B has the same dimension. This quantum information source is therefore a "conditional quantum information source." Let $\mathcal{H}_B$ and B denote the respective Hilbert space and system label corresponding to the quantum output of the conditional quantum information source. Let us call the resulting ensemble the "classical-quantum ensemble" and say that a "classical-quantum information source" generates it. The classical-quantum ensemble is as follows:

$$\left\{p_X(x), |x\rangle\langle x|^X \otimes \rho_x^B\right\}_{x\in\mathcal{X}}, \tag{14.48}$$

where we correlate the classical state $|x\rangle^X$ with the density operator ρ_x^B of the conditional quantum information source. The expected density operator of the above classical-quantum ensemble is the following classical-quantum state (discussed in Section 4.3.4):

$$\rho^{XB} \equiv \sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x|^X \otimes \rho_x^B. \tag{14.49}$$

The conditional quantum entropy $H(B|X)$ of the classical-quantum state ρ^{XB} is as follows:

$$H(B|X) = \sum_{x\in\mathcal{X}} p_X(x) H(B|X=x) = \sum_{x\in\mathcal{X}} p_X(x) H\big(\rho_x^B\big). \tag{14.50}$$

We can write the spectral decomposition of each conditional density operator ρ_x^B as follows:

$$\sum_{y\in\mathcal{Y}} p_{Y|X}(y|x)|y_x\rangle\langle y_x|^B, \tag{14.51}$$

where the elements of the set $\{y\}_{y\in\mathcal{Y}}$ label the elements of an alphabet $\mathcal{Y}$, the orthonormal set $\{|y_x\rangle^B\}_{y\in\mathcal{Y}}$ is the set of eigenvectors of ρ_x^B, and the corresponding eigenvalues are $\big\{p_{Y|X}(y|x)\big\}_{y\in\mathcal{Y}}$. We need the x label for the orthonormal set $\{|y_x\rangle^B\}_{y\in\mathcal{Y}}$ because the decomposition may be different for different density operators ρ_x^B. The above notation is again suggestive because the eigenvalues $p_{Y|X}(y|x)$ correspond to conditional probabilities, and the set $\{|y_x\rangle^B\}$ of eigenvectors corresponds to an orthonormal set of quantum states conditioned

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

on label x. With this respresentation, the conditional entropy $H(B|X)$ reduces to a formula that looks like that for the classical conditional entropy:

$$H(B|X) = \sum_{x\in\mathcal{X}} p_X(x) H\left(\rho_x^B\right) \tag{14.52}$$

$$= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_X(x) p_{Y|X}(y|x) \log \frac{1}{p_{Y|X}(y|x)}. \tag{14.53}$$

We now consider when the classical-quantum information source emits a large number n of states. The density operator for the output state $\rho^{X^n B^n}$ is as follows:

$$\rho^{X^n B^n}$$

$$\equiv \left(\rho^{XB}\right)^{\otimes n} \tag{14.54}$$

$$= \left(\sum_{x_1\in\mathcal{X}} p_X(x_1)|x_1\rangle\langle x_1|^{X_1} \otimes \rho_{x_1}^{B_1}\right) \otimes \cdots \otimes \left(\sum_{x_n\in\mathcal{X}} p_X(x_n)|x_n\rangle\langle x_n|^{X_n} \otimes \rho_{x_n}^{B_n}\right) \tag{14.55}$$

$$= \sum_{x_1,\ldots,x_n\in\mathcal{X}} p_X(x_1)\cdots p_X(x_n)|x_1\rangle\cdots|x_n\rangle\langle x_1|\cdots\langle x_n|^{X^n} \otimes \left(\rho_{x_1}^{B_1}\otimes\cdots\otimes\rho_{x_n}^{B_n}\right). \tag{14.56}$$

We can abbreviate the above state as

$$\sum_{x^n\in\mathcal{X}^n} p_{X^n}(x^n)|x^n\rangle\langle x^n|^{X^n} \otimes \rho_{x^n}^{B^n}, \tag{14.57}$$

where

$$p_{X^n}(x^n) \equiv p_X(x_1)\cdots p_X(x_n), \tag{14.58}$$

$$|x^n\rangle^{X^n} \equiv |x_1\rangle^{X_1}\cdots|x_n\rangle^{X_n}, \qquad \rho_{x^n}^{B^n} \equiv \rho_{x_1}^{B_1}\otimes\cdots\otimes\rho_{x_n}^{B_n}, \tag{14.59}$$

and the spectral decomposition for the state $\rho_{x^n}^{B^n}$ is

$$\rho_{x^n}^{B^n} = \sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|x^n)|y_{x^n}^n\rangle\langle y_{x^n}^n|^{B^n}, \tag{14.60}$$

where

$$p_{Y^n|X^n}(y^n|x^n) \equiv p_{Y_1|X_1}(y_1|x_1)\cdots p_{Y_n|X_n}(y_n|x_n), \tag{14.61}$$

$$|y_{x^n}^n\rangle^{B^n} \equiv |\,y_{1\,x_1}\rangle^{B_1}\cdots|\,y_{n\,x_n}\rangle^{B_n}. \tag{14.62}$$

The above developments are a step along the way for formulating the definitions of weak and strong conditional quantum typicality.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

14.2.1 Weak Conditional Quantum Typicality

We can "quantize" the notion of weak classical conditional typicality so that it applies to a classical-quantum information source.

Definition 14.2.1 (Weak Conditionally Typical Subspace)**.** *The conditionally typical subspace $T_\delta^{B^n|x^n}$ corresponds to a particular sequence x^n and an ensemble $\{p_X(x), \rho_x^B\}$. It is the subspace spanned by the states $|y_{x^n}^n\rangle^{B^n}$ whose conditional sample entropy is δ-close to the true conditional quantum entropy:*

$$T_\delta^{B^n|x^n} \equiv \text{span}\Big\{|y_{x^n}^n\rangle^{B^n} : \big|\overline{H}(y^n|x^n) - H(B|X)\big| \leq \delta\Big\}, \tag{14.63}$$

where the states $|y_{x^n}^n\rangle^{B^n}$ are formed from the eigenstates of the density operators ρ_x^B (they are of the form in (14.62)) and the sample entropy is with respect to the distribution $p_{Y|X}(y|x)$ from (14.51).

Definition 14.2.2 (Weak Conditionally Typical Projector)**.** *The projector $\Pi_\delta^{B^n|x^n}$ onto the conditionally typical subspace $T_\delta^{B^n|x^n}$ is as follows:*

$$\Pi_\delta^{B^n|x^n} \equiv \sum_{|y_{x^n}^n\rangle \in T_\delta^{B^n|x^n}} |y_{x^n}^n\rangle\langle y_{x^n}^n|^{B^n}. \tag{14.64}$$

14.2.2 Properties of the Weak Conditionally Typical Subspace

The weak conditionally typical subspace $T_\delta^{B^n|x^n}$ enjoys several useful properties that are "quantized" versions of the properties for weak conditionally typical sequences discussed in Section 13.6. We should point out that we cannot really say much for several of the properties for a particular sequence x^n, but we can do so on average for a random sequence X^n. Thus, several of the properties give expected behavior for a random sequence X^n. This convention for quantum weak conditional typicality is the same as we had for classical weak conditional typicality in Section 13.6.

Property 14.2.1 (Unit Probability) The expectation of the probability that we measure a random quantum state $\rho_{X^n}^{B^n}$ to be in the conditionally typical subspace $T_\delta^{B^n|X^n}$ approaches one as n becomes large:

$$\forall \epsilon > 0 \qquad \mathbb{E}_{X^n}\Big\{\text{Tr}\Big\{\Pi_\delta^{B^n|X^n}\rho_{X^n}^{B^n}\Big\}\Big\} \geq 1 - \epsilon \quad \text{for sufficiently large } n. \tag{14.65}$$

Property 14.2.2 (Exponentially Small Dimension) The dimension $\dim(T_\delta^{B^n|x^n})$ of the δ-conditionally typical subspace is exponentially smaller than the dimension $|\mathcal{Y}|^n$ of the entire space of quantum states for most classical-quantum sources. We formally state this property as follows:

$$\text{Tr}\Big\{\Pi_\delta^{B^n|x^n}\Big\} \leq 2^{n(H(B|X)+\delta)}. \tag{14.66}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can also lower bound the dimension $\dim(T_\delta^{B^n|x^n})$ of the δ-conditionally typical subspace when n is sufficiently large:

$$\forall \epsilon > 0 \qquad \mathbb{E}_{X^n}\left\{ \operatorname{Tr}\left\{ \Pi_\delta^{B^n|X^n} \right\} \right\} \geq (1-\epsilon) 2^{n(H(B|X)-\delta)} \quad \text{for sufficiently large } n. \tag{14.67}$$

Property 14.2.3 (Equipartition) The density operator $\rho_{x^n}^{B^n}$ looks approximately maximally mixed when projected to the conditionally typical subspace:

$$2^{-n(H(B|X)+\delta)} \Pi_\delta^{B^n|x^n} \leq \Pi_\delta^{B^n|x^n} \rho_{x^n}^{B^n} \Pi_\delta^{B^n|x^n} \leq 2^{-n(H(B|X)-\delta)} \Pi_\delta^{B^n|x^n}. \tag{14.68}$$

Exercise 14.2.1 Prove all three of the above properties for weak conditional quantum typicality.

14.2.3 Strong Conditional Quantum Typicality

We now develop the notion of strong conditional quantum typicality. This notion again applies to an ensemble or to a classical-quantum state such as that given in (14.49). Though, it differs from weak conditional quantum typicality because we can prove stronger statements about the asymptotic behavior of conditional quantum systems (just as we could for the classical case in Section 13.9). We begin this section with an example to build up our intuition. We then follow with the formal definition of strong conditional quantum typicality, and we end by proving some properties of the strong conditionally typical subspace.

Recall the example from Section 13.7. In a similar way to this example, we can draw a sequence from an alphabet $\{0, 1, 2\}$ according to the following distribution:

$$p_X(0) = \frac{1}{4}, \qquad p_X(1) = \frac{1}{4}, \qquad p_X(2) = \frac{1}{2}. \tag{14.69}$$

One potential realization sequence is as follows:

$$201020102212. \tag{14.70}$$

The above sequence has four "zeros," three "ones," and five "twos," so that the empirical distribution of this sequence is $(1/3, 1/4, 5/12)$ and has maximum deviation $1/12$ from the true distribution in (14.69).

For each symbol in the above sequence, we could then draw from one of three quantum information sources based on whether the classical index is 0, 1, or 2. Suppose that the expected density operator of the first quantum information source is ρ_0, that of the second is ρ_1, and that of the third is ρ_2. Then the density operator for the resulting sequence of quantum states is as follows:

$$\rho_2^{B_1} \otimes \rho_0^{B_2} \otimes \rho_1^{B_3} \otimes \rho_0^{B_4} \otimes \rho_2^{B_5} \otimes \rho_0^{B_6} \otimes \rho_1^{B_7} \otimes \rho_0^{B_8} \otimes \rho_2^{B_9} \otimes \rho_2^{B_{10}} \otimes \rho_1^{B_{11}} \otimes \rho_2^{B_{12}}, \tag{14.71}$$

where the superscripts label the quantum systems in which the density operators live. So, the state of systems B_1, B_5, B_9, B_{10}, and B_{12} is equal to five copies of ρ_2, the state of systems

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

B_2, B_4, B_6, and B_8 is equal to four copies of ρ_0, and the state of systems B_3, B_7, and B_{11} is equal to three copies of ρ_1. Let I_x be an indicator set for each $x \in \{0,1,2\}$, so that I_x consists of all the indices in the sequence for which a symbol is equal to x. For the above example,

$$I_0 = \{2,4,6,8\}, \qquad I_1 = \{3,7,11\}, \qquad I_2 = \{1,5,9,10,12\}. \tag{14.72}$$

These sets serve as a way of grouping all of the density operators that are the same because they correspond to the same classical symbol, and it is important to do so if we would like to consider concentration of measure effects when we go to the asymptotic setting. As a visual aid, we could permute the sequence of density operators in (14.71) if we would like to see systems with the same density operator grouped together:

$$\rho_0^{B_2} \otimes \rho_0^{B_4} \otimes \rho_0^{B_6} \otimes \rho_0^{B_8} \otimes \rho_1^{B_3} \otimes \rho_1^{B_7} \otimes \rho_1^{B_{11}} \otimes \rho_2^{B_1} \otimes \rho_2^{B_5} \otimes \rho_2^{B_9} \otimes \rho_2^{B_{10}} \otimes \rho_2^{B_{12}}. \tag{14.73}$$

There is then a typical projector for the first four systems with density operator ρ_0, a different typical projector for the next three systems with density operator ρ_1, and an even different typical projector for the last five systems with density operator ρ_2 (though, the length of the above quantum sequence is certainly not large enough to observe any measure concentration effects!). Thus, the indicator sets I_x serve to identify which systems have the same density operator so that we can know upon which systems a particular typical projector should act.

This example helps build our intuition of strong conditional quantum typicality, and we can now begin to state what we would expect in the asymptotic setting. Suppose that the original classical sequence is large and strongly typical, so that it has roughly $n/4$ occurrences of "zero," $n/4$ occurrences of "one," and $n/2$ occurrences of "two." We would then expect the law of large numbers to come into play for $n/4$ and $n/2$ when n is large enough. Thus, we can use the classical sequence to identify which quantum systems have the same density operator, and apply a typical projector to each of these subsets of quantum systems. Then all of the useful asymptotic properties of typical subspaces apply whenever n is large enough.

We can now state the definition of the strong conditionally typical subspace and the strong conditionally typical projector, and we prove some of their asymptotic properties by exploiting the properties of typical subspaces.

Definition 14.2.3 (Strong Conditionally Typical Subspace)**.** *The strong conditionally typical subspace corresponds to a sequence x^n and an ensemble $\{p_X(x), \rho_x^B\}$. Let the spectral decomposition of each state ρ_x^B be as in (14.51) with distribution $p_{Y|X}(y|x)$ and corresponding eigenstates $|y_x\rangle$. The strong conditionally typical subspace $T_\delta^{B^n|x^n}$ is then as follows:*

$$T_\delta^{B^n|x^n} \equiv \text{span}\left\{ \bigotimes_{x \in \mathcal{X}} |y_x^{I_x}\rangle^{B^{I_x}} : \forall x, \quad y^{I_x} \in T_\delta^{(Y|x)^{|I_x|}} \right\}, \tag{14.74}$$

where $I_x \equiv \{i : x_i = x\}$ is an indicator set that selects the indices i in the sequence x^n for which the i^{th} symbol x_i is equal to $x \in \mathcal{X}$, B^{I_x} selects the systems from B^n where the classical sequence x^n is equal to the symbol x, $|y_x^{I_x}\rangle$ is some string of states from the set $\{|y_x\rangle\}$, y^{I_x} is a classical string corresponding to this string of states, $Y|x$ is a random variable with distribution $p_{Y|X}(y|x)$, and $|I_x|$ is the cardinality of the indicator set I_x.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Definition 14.2.4 (Strong Conditionally Typical Projector)**.** *The strong conditionally typical projector again corresponds to a sequence x^n and an ensemble $\{p_X(x), \rho_x^B\}$. It is a tensor product of typical projectors for each state ρ_x^B in the ensemble:*

$$\Pi_\delta^{B^n|x^n} \equiv \bigotimes_{x\in\mathcal{X}} \Pi_{\rho_x,\delta}^{B^{I_x}}, \tag{14.75}$$

where I_x is defined in Definition 14.2.3, and B^{I_x} indicates the systems onto which a particular typical projector for ρ_x projects.[2]

14.2.4 Properties of the Strong Conditionally Typical Subspace

The strong conditionally typical subspace admits several useful asymptotic properties similar to what we have seen before, and the proof strategy for proving all of them is similar to the way that we proved the analogous properties for the strong conditionally typical set in Section 13.9.3. Suppose that we draw a sequence x^n from a probability distribution $p_X(x)$, and we are able to draw as many samples as we wish so that the sequence x^n is strongly typical and the occurrences $N(x|x^n)$ of each symbol x are as large as as we wish. Then the following properties hold.

Property 14.2.4 (Unit Probability) The probability that we measure a quantum state $\rho_{x^n}^{B^n}$ to be in the conditionally typical subspace $T_\delta^{B^n|x^n}$ approaches one as n becomes large:

$$\forall \epsilon > 0 \qquad \text{Tr}\left\{\Pi_\delta^{B^n|x^n} \rho_{x^n}^{B^n}\right\} \geq 1 - \epsilon \quad \text{for sufficiently large } n. \tag{14.76}$$

Property 14.2.5 (Exponentially Small Dimension) The dimension $\dim(T_\delta^{B^n|x^n})$ of the δ-conditionally typical subspace is exponentially smaller than the dimension $|B|^n$ of the entire space of quantum states for all classical-quantum information sources besides ones where all their density operators are maximally mixed. We formally state this property as follows:

$$\text{Tr}\left\{\Pi_\delta^{B^n|x^n}\right\} \leq 2^{n(H(B|X)+\delta'')}. \tag{14.77}$$

We can also lower bound the dimension $\dim(T_\delta^{Y^n|x^n})$ of the δ-conditionally typical subspace when n is sufficiently large:

$$\forall \epsilon > 0 \qquad \text{Tr}\left\{\Pi_\delta^{B^n|x^n}\right\} \geq (1-\epsilon)2^{n(H(B|X)-\delta'')} \quad \text{for sufficiently large } n. \tag{14.78}$$

Property 14.2.6 (Equipartition) The state $\rho_{x^n}^{B^n}$ is approximately maximally mixed when projected onto the strong conditionally typical subspace:

$$2^{-n(H(B|X)+\delta'')}\Pi_\delta^{B^n|x^n} \leq \Pi_\delta^{B^n|x^n} \rho_{x^n}^{B^n} \Pi_\delta^{B^n|x^n} \leq 2^{-n(H(B|X)-\delta'')}\Pi_\delta^{B^n|x^n}. \tag{14.79}$$

[2]Having the conditional density operators in the subscript breaks somewhat from our convention throughout this chapter, but it is useful here to indicate explicitly which density operator corresponds to a typical projector.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

14.2.5 Proofs of the Properties of the Strong Conditionally Typical Subspace

Proof of the Unit Probability Property (Property 14.2.4). The proof of this property is similar to the proof of Property 13.9.1 for the strong conditionally typical set. Since we are dealing with an IID distribution, we can assume without loss of generality that the sequence x^n is lexicographically ordered with an order on the alphabet $\mathcal{X}$. We write the elements of $\mathcal{X}$ as $a_1, \ldots, a_{|\mathcal{X}|}$. Then the lexicographic ordering means that we can write the sequence of quantum states ρ_{x^n} as follows:

$$\rho_{x^n} = \underbrace{\rho_{a_1} \otimes \cdots \otimes \rho_{a_1}}_{N(a_1|x^n)} \otimes \underbrace{\rho_{a_2} \otimes \cdots \otimes \rho_{a_2}}_{N(a_2|x^n)} \otimes \cdots \otimes \underbrace{\rho_{a_{|\mathcal{X}|}} \otimes \cdots \otimes \rho_{a_{|\mathcal{X}|}}}_{N\left(a_{|\mathcal{X}|}|x^n\right)}. \tag{14.80}$$

It follows that $N(a_i|x^n) \geq n(p_X(a_i) - \delta')$ from the typicality of x^n, and thus the law of large numbers comes into play for each block $a_i \cdots a_i$ with length $N(a_i|x^n)$. The strong conditionally typical projector $\Pi_\delta^{B^n|x^n}$ for this system is as follows:

$$\Pi_\delta^{B^n|x^n} \equiv \bigotimes_{x \in \mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}}, \tag{14.81}$$

because we assumed the lexicographic ordering of the symbols in the sequence x^n. Each projector $\Pi_{\rho_x,\delta}^{B^{N(x|x^n)}}$ in the above tensor product is a typical projector for the density operator ρ_x when $N(x|x^n) \approx np_X(x)$ becomes very large. Then we can apply the Unit Probability Property (Property 14.1.1) for each of these typical projectors, and it follows that

$$\begin{aligned}
\operatorname{Tr}\left\{\rho_{x^n}^{B^n} \Pi_\delta^{B^n|x^n}\right\} &= \operatorname{Tr}\left\{\bigotimes_{x \in \mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} \rho_x^{\otimes N(x|x^n)}\right\} && (14.82)\\
&= \prod_{x \in \mathcal{X}} \operatorname{Tr}\left\{\Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} \rho_x^{\otimes N(x|x^n)}\right\} && (14.83)\\
&\geq (1-\epsilon)^{|\mathcal{X}|} && (14.84)\\
&\geq 1 - |\mathcal{X}|\epsilon. && (14.85)
\end{aligned}$$

□

Proof of the Equipartition Property (Property 14.2.6). We first assume without loss of generality that we can write the state $\rho_{x^n}^{B^n}$ in lexicographic order as in (14.80). Then the strong conditionally typical projector is again as in (14.81). It follows that

$$\Pi_\delta^{B^n|x^n} \rho_{x^n} \Pi_\delta^{B^n|x^n} = \bigotimes_{x \in \mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} \rho_x^{\otimes N(x|x^n)} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}}. \tag{14.86}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can apply the Equipartition Property of the typical subspace for each typical projector $\Pi_{\rho_x,\delta}^{B^{N(x|x^n)}}$ (Property 14.1.3):

$$\bigotimes_{x\in\mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} 2^{-N(x|x^n)(H(\rho_x)+c\delta)} \leq \bigotimes_{x\in\mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} \rho_x^{\otimes N(x|x^n)} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} \leq \bigotimes_{x\in\mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} 2^{-N(x|x^n)(H(\rho_x)-c\delta)} \quad (14.87)$$

The following inequalities hold because the sequence x^n is strongly typical as defined in Definition 13.7.2:

$$\bigotimes_{x\in\mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} 2^{-n(p_X(x)+\delta')(H(\rho_x)+c\delta)} \leq \Pi_\delta^{B^n|x^n} \rho_{x^n} \Pi_\delta^{B^n|x^n} \leq \bigotimes_{x\in\mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} 2^{-n(p_X(x)-\delta')(H(\rho_x)-c\delta)}. \quad (14.88)$$

We can factor out each term $2^{-n(p_X(x)+\delta')(H(\rho_x)+c\delta)}$ from the tensor products:

$$\prod_{x\in\mathcal{X}} 2^{-n(p_X(x)+\delta')(H(\rho_x)+c\delta)} \bigotimes_{x\in\mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}} \leq \Pi_\delta^{B^n|x^n} \rho_{x^n} \Pi_\delta^{B^n|x^n} \leq \prod_{x\in\mathcal{X}} 2^{-n(p_X(x)-\delta')(H(\rho_x)-c\delta)} \bigotimes_{x\in\mathcal{X}} \Pi_{\rho_x,\delta}^{B^{N(x|x^n)}}. \quad (14.89)$$

We then multiply out the $|\mathcal{X}|$ terms $2^{-n(p_X(x)+\delta')(H(\rho_x)+c\delta)}$:

$$2^{-n\left(H(B|X)+\sum_x (H(\rho_x)\delta'+cp_X(x)\delta+c\delta\delta')\right)} \Pi_\delta^{B^n|x^n} \leq \Pi_\delta^{B^n|x^n} \rho_{x^n} \Pi_\delta^{B^n|x^n} \leq 2^{-n\left(H(B|X)+\sum_x c\delta\delta'-H(\rho_x)\delta'-cp_X(x)\delta\right)} \Pi_\delta^{B^n|x^n}. \quad (14.90)$$

The final step below follows because $\sum_x p_X(x) = 1$ and because the bound $\sum H(\rho_x) \leq |\mathcal{X}| \log d$ applies where d is the dimension of the density operator ρ_x:

$$2^{-n(H(B|X)+\delta'')} \Pi_\delta^{B^n|x^n} \leq \Pi_\delta^{B^n|x^n} \rho_{x^n} \Pi_\delta^{B^n|x^n} \leq 2^{-n(H(B|X)-\delta'')} \Pi_\delta^{B^n|x^n}, \quad (14.91)$$

where

$$\delta'' \equiv \delta' |\mathcal{X}| \log d + c\delta + |\mathcal{X}| c\delta\delta'. \quad (14.92)$$

□

Exercise 14.2.2 Prove Property 14.2.5.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

14.2.6 Strong Conditional and Marginal Quantum Typicality

We end this section on strong conditional quantum typicality by proving a final property that applies to a state drawn from an ensemble and the typical subspace of the expected density operator of the ensemble.

Property 14.2.7 Consider an ensemble of the form $\{p_X(x), \rho_x\}$ with expected density operator $\rho \equiv \sum_x p_X(x)\rho_x$. Suppose that x^n is a strongly typical sequence with respect to the distribution $p_X(x)$ and leads to a conditional density operator ρ_{x^n}. Then the probability of measuring ρ_{x^n} in the strongly typical subspace of ρ is high:

$$\forall \epsilon > 0 \qquad \text{Tr}\{\Pi^n_{\rho,\delta}\ \rho_{x^n}\} \geq 1 - \epsilon, \qquad \text{for sufficiently large } n, \tag{14.93}$$

where the typical projector $\Pi^n_{\rho,\delta}$ is with respect to the density operator ρ.

Proof. Let the expected density operator have the following spectral decomposition:

$$\rho = \sum_z p_Z(z)|z\rangle\langle z|. \tag{14.94}$$

We define the "pinching" operation as a dephasing with respect to the basis $\{|z\rangle\}$:

$$\sigma \to \Delta(\sigma) \equiv \sum_z |z\rangle\langle z|\sigma|z\rangle\langle z|. \tag{14.95}$$

Let $\overline{\rho}_x$ denote the pinched version of the conditional density operators ρ_x:

$$\overline{\rho}_x \equiv \Delta(\rho_x) = \sum_z |z\rangle\langle z|\rho_x|z\rangle\langle z| = \sum_z p_{Z|X}(z|x)|z\rangle\langle z|, \tag{14.96}$$

where $p_{Z|X}(z|x) \equiv \langle z|\rho_x|z\rangle$. This pinching is the crucial insight for the proof because all of the pinched density operators $\overline{\rho}_x$ have a common eigenbasis and the analysis reduces from a quantum one to a classical one that exploits the properties of strong marginal, conditional, and joint typicality. The following chain of inequalities then holds by exploiting the above

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

definitions:

$$\operatorname{Tr}\{\rho_{x^n}\Pi^n_{\rho,\delta}\} = \operatorname{Tr}\left\{\rho_{x^n}\sum_{z^n\in T^{Z^n}_\delta}|z^n\rangle\langle z^n|\right\} \tag{14.97}$$

$$= \operatorname{Tr}\left\{\rho_{x^n}\sum_{z^n\in T^{Z^n}_\delta}|z^n\rangle\langle z^n|z^n\rangle\langle z^n|\right\} \tag{14.98}$$

$$= \operatorname{Tr}\left\{\sum_{z^n\in T^{Z^n}_\delta}|z^n\rangle\langle z^n|\rho_{x^n}|z^n\rangle\langle z^n|\right\} \tag{14.99}$$

$$= \operatorname{Tr}\left\{\sum_{z^n\in T^{Z^n}_\delta}p_{Z^n|X^n}(z^n|x^n)|z^n\rangle\langle z^n|\right\} \tag{14.100}$$

$$= \sum_{z^n\in T^{Z^n}_\delta}p_{Z^n|X^n}(z^n|x^n) \tag{14.101}$$

The first equality follows from the definition of the typical projector $\Pi^n_{\rho,\delta}$. The second equality follows because $|z^n\rangle\langle z^n|$ is a projector, and the third follows from linearity and cyclicity of the trace. The fourth equality follows because

$$\langle z^n|\rho_{x^n}|z^n\rangle = \prod_{i=1}^n\langle z_i|\rho_{x_i}|z_i\rangle = \prod_{i=1}^n p_{Z|X}(z_i|x_i) \equiv p_{Z^n|X^n}(z^n|x^n). \tag{14.102}$$

Now consider this final expression $\sum_{z^n\in T^{Z^n}_\delta}p_{Z^n|X^n}(z^n|x^n)$. It is equivalent to the probability that a random conditional sequence $Z^n|x^n$ is in the typical set for $p_Z(z)$:

$$\Pr\{Z^n|x^n \in T^{Z^n}_\delta\}. \tag{14.103}$$

By taking n large enough, the law of large numbers guarantees that it is highly likely (with probability greater than $1-\epsilon$ for any $\epsilon > 0$) that this random conditional sequence $Z^n|x^n$ is in the conditionally typical set $T^{Z^n|x^n}_{\delta'}$ for some δ'. It then follows that this conditional sequence has a high probability of being in the unconditionally typical set $T^{Z^n}_\delta$ because we assumed that the sequence x^n is strongly typical and Lemma 13.9.1 states that a sequence z^n is unconditionally typical if x^n is strongly typical and z^n is strong conditionally typical. □

14.3 The Method of Types for Quantum Systems

Our final development in this chapter is to establish the method of types in the quantum domain, and the classical tools from Section 13.7 have a straightforward generalization.

We can partition the Hilbert space of n qudits into different type class subspaces, just as we can partition the set of all sequences into different type classes. For example, consider

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

the Hilbert space of three qubits. The computational basis is an orthonormal basis for the entire Hilbert space of three qubits:

$$\{|000\rangle, |001\rangle, |010\rangle, |011\rangle, |100\rangle, |101\rangle, |110\rangle, |111\rangle\}. \tag{14.104}$$

Then the computational basis states with the same Hamming weight form a basis for each type class subspace. So, for the above example, the type class subspaces are as follows:

$$T_0 \equiv \{|000\rangle\}, \tag{14.105}$$
$$T_1 \equiv \{|001\rangle, |010\rangle, |100\rangle\}, \tag{14.106}$$
$$T_2 \equiv \{|011\rangle, |101\rangle, |110\rangle\}, \tag{14.107}$$
$$T_3 \equiv \{|111\rangle\}, \tag{14.108}$$

and the projectors onto the different type class subspaces are as follows:

$$\Pi_0 \equiv |000\rangle\langle 000|, \tag{14.109}$$
$$\Pi_1 \equiv |001\rangle\langle 001| + |010\rangle\langle 010| + |100\rangle\langle 100|, \tag{14.110}$$
$$\Pi_2 \equiv |011\rangle\langle 011| + |101\rangle\langle 101| + |110\rangle\langle 110|, \tag{14.111}$$
$$\Pi_3 \equiv |111\rangle\langle 111|. \tag{14.112}$$

We can generalize the above example to an n-fold tensor product of qudit systems with the method of types.

Definition 14.3.1 (Type Class Subspace)**.** *The type class subspace is the subspace spanned by all states with the same type:*

$$T_t^{X^n} \equiv \text{span}\left\{|x^n\rangle : x^n \in T_t^{X^n}\right\}, \tag{14.113}$$

where the notation T_t^n on the LHS indicates the type class subspace, and the notation $T_t^{X^n}$ on the RHS indicates the type class of the classical sequence x^n.

Definition 14.3.2 (Type Class Projector)**.** *Let $\Pi_t^{X^n}$ denote the type class subspace projector:*

$$\Pi_t^n \equiv \sum_{x^n \in T_t^{X^n}} |x^n\rangle\langle x^n|. \tag{14.114}$$

Property 14.3.1 (Resolution of the Identity with Type Class Projectors) The sum of all type class projectors forms a resolution of the identity on the full Hilbert space $\mathcal{H}^{\otimes n}$ of n qudits:

$$I = \sum_t \Pi_t^n, \tag{14.115}$$

where I is the identity operator on $\mathcal{H}^{\otimes n}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Definition 14.3.3 (Maximally Mixed Type Class State)**.** *The maximally mixed density operator proportional to the type class subspace projector is*

$$\pi_t \equiv {D_t}^{-1}\Pi_t^{X^n}, \tag{14.116}$$

where D_t is the dimension of the type class:

$$D_t \equiv Tr\{\Pi_t^{X^n}\}. \tag{14.117}$$

Recall from Definition 13.7.4 that a δ-typical type is one for which the empirical distribution has maximum deviation δ from the true distribution, and τ_δ is the set of all δ-typical types. For the quantum case, we determine the maximum deviation δ of a type from the true distribution $p_X(x)$ (this is the distribution from the spectral decomposition of a density operator ρ). This definition allows us to write the strongly δ-typical subspace projector $\Pi_\delta^{X^n}$ of ρ as a sum over all of the δ-typical type class projectors $\Pi_t^{X^n}$:

$$\Pi_\delta^{X^n} = \sum_{t\in\tau_\delta} \Pi_t^{X^n}. \tag{14.118}$$

Some protocols in quantum Shannon theory such as entanglement concentration in Chapter 18 employ the above decomposition of the typical subspace projector into types. The way that such a protocol works is first to perform a typical subspace measurement on many copies of a state, and this measurement succeeds with high probability. One party involved in the protocol then performs a type class measurement $\{\Pi_t^{X^n}\}_t$. We perform this latter measurement in a protocol if we would like the state to have a uniform distribution over states in the type class. One might initially think that the dimension of the remaining state would not be particularly large, but it actually holds that the dimension is large because we can obtain the following useful lower bound on the dimension of any typical type class projector.

Property 14.3.2 (Minimal Dimension of a Typical Type Class Projector) Suppose that $p_X(x)$ is the distribution from the spectral decomposition of a density operator ρ, and τ_δ collects all the type class subspaces with maximum deviation δ from the distribution $p_X(x)$. Then for any type $t \in \tau_\delta$ and for sufficiently large n, we can lower bound the dimension of the type class projector $\Pi_t^{X^n}$ as follows:

$$\mathrm{Tr}\{\Pi_t^{X^n}\} \geq 2^{n[H(\rho)-\eta(d\delta)-d\frac{1}{n}\log(n+1)]}, \tag{14.119}$$

where d is the dimension of the Hilbert space where ρ lives and the function $\eta(d\delta) \to 0$ as $\delta \to 0$.

Proof. The proof follows directly by exploiting Property 13.7.5 from the previous chapter. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

14.4 Concluding Remarks

This chapter is about the asymptotic nature of quantum information in the IID setting. The main technical development is the notion of the typical subspace, and our approach here is simply to "quantize" the definition of the typical set from the previous chapter. The typical subspace enjoys properties similar to those of the typical set—the probability that many copies of a density operator lie in the typical subspace approaches one as the number of copies approaches infinity, the dimension of the typical subspace is exponentially smaller than the dimension of the full Hilbert space, and many copies of a density operator look approximately maximally mixed on the typical subspace. The rest of the content in this chapter involves an extension of these ideas to conditional quantum typicality.

The content in this chapter is here to provide a rigorous underpinning that we can quickly cite later on, and after having mastered the results in this chapter along with the tools in the next two chapters, we will be ready to prove many of the important results in quantum Shannon theory.

14.5 History and Further Reading

Ohya and Petz devised the notion of a typical subspace [200], and later Schumacher independently devised it when he proved the quantum data compression theorem bearing his name [216]. Holevo [144], Schumacher, and Westmoreland [219] introduced the conditionally typical subspace in order to prove the HSW coding theorem. Winter's thesis is a good source for proofs of several properties of quantum typicality [256]. The book of Nielsen and Chuang uses weak conditional quantum typicality to prove the HSW theorem [197]. Bennett *et al.* [34] and Holevo [146] introduced frequency-typical (or strongly-typical) subspaces to quantum information theory in order to prove the entanglement-assisted classical capacity theorem. Devetak used strong typicality to prove the HSW coding theorem in Appendix B of Ref. [68].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 15

The Packing Lemma

The Packing Lemma is a general method for one party to pack classical messages into a Hilbert space so that another party can distinguish the packed messages. The first party has access to an ensemble of quantum states, and the other party has access to a set of projectors from which he can form a quantum measurement. If the ensemble and the projectors satisfy the conditions of the Packing Lemma, then it guarantees the existence of scheme by which the second party can distinguish the classical messages that the first party prepares.

The statement of the Packing Lemma is quite general, and this approach has a great advantage because we can use it as a primitive in many coding theorems in quantum Shannon theory. Examples of coding theorems that we can prove with the Packing Lemma are the Holevo-Schumacher-Westmoreland (HSW) theorem for transmission of classical information over a quantum channel and the entanglement-assisted classical capacity theorem for the transmission of classical information over an entanglement-assisted quantum channel (furthermore, Chapter 21 shows that these two protocols are sufficient to generate most known protocols in quantum Shannon theory). Combined with the Covering Lemma of the next chapter, the Packing Lemma gives a method for transmitting private classical information over a quantum channel, and this technique in turn gives a way to communicate quantum information over a quantum channel. As long as we can determine an ensemble and a set of projectors satisfying the conditions of the Packing Lemma, we can apply it in a straightforward way. For example, we prove the HSW coding theorem in Chapter 19 largely by relying on the properties of typical and conditionally typical subspaces that we proved in the previous chapter, and some of these properties are equivalent to the conditions of the Packing Lemma.

The Packing Lemma is a "one-shot" lemma because it applies to a general scenario that is not limited only to IID uses of a quantum channel. This "one-shot" approach is part of the reason that we can apply it in so many different situations. The technique of proving a "one-shot" result and applying it to the IID scenario is a common method of attack in quantum Shannon theory (we do it again in Chapter 16 by proving the Covering Lemma that helps in determining a way to send private classical information over a noisy quantum channel).

We begin in the next section with a simple example that illustrates the main ideas of the Packing Lemma. We then generalize this setting and give the statement of the Packing Lemma. We dissect its proof in several sections that explain the random selection of a code, the construction of a quantum measurement, and the error analysis. We finally show how to derandomize the Packing Lemma so that there exists some scheme for packing classical messages into Hilbert space with negligible probability of error for determining each classical message.

15.1 Introductory Example

Suppose that Alice would like to communicate classical information to Bob, and suppose further that she can prepare a message for Bob using the following BB84 ensemble:

$$\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}, \tag{15.1}$$

where each state occurs with equal probability. Let us label each of the above states by the classical indices a, b, c, and d so that a labels $|0\rangle$, b labels $|1\rangle$, etc. She cannot use all of the states for transmitting classical information because, for example, $|0\rangle$ and $|+\rangle$ are non-orthogonal states and there is no measurement that can distinguish them with high probability.

How can Alice communicate to Bob using this ensemble? She can choose a subset of the states in the BB84 ensemble for transmitting classical information. She can choose the states $|0\rangle$ and $|1\rangle$ for encoding one classical bit of information. Bob can then perform a Von Neumann measurement in the basis $\{|0\rangle, |1\rangle\}$ to determine the message that Alice encodes. Alternatively, Alice and Bob can use the states $|+\rangle$ and $|-\rangle$ in a similar fashion for encoding one classical bit of information.

In the above example, Alice can send two messages by using the labels a and b only. We say that the labels a and b constitute the *code*. The states $|0\rangle$ and $|1\rangle$ are the *codewords*, the projectors $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$ are each a *codeword projector*, and the projector $|0\rangle\langle 0| + |1\rangle\langle 1|$ is the *code projector* (in this case, the code projector projects onto the whole Hilbert space).

The construction in the above example gives a way to use a certain ensemble for "packing" classical information into Hilbert space, but there is only so much room for packing. For example, it is impossible to encode more than one bit of classical information into a qubit such that someone else can access this classical information reliably—this is the statement of the Holevo bound (Exercise 11.9.1).

15.2 The Setting of the Packing Lemma

We generalize the above example to show how Alice can efficiently pack classical information into a Hilbert space such that Bob can retrieve it with high probability. Suppose that Alice's resource for communication is an ensemble $\{p_X(x), \sigma_x\}_{x\in\mathcal{X}}$ of quantum states that she can

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

prepare for Bob, where the states σ_x are not necessarily perfectly distinguishable. We define the ensemble as follows:

Definition 15.2.1 (Ensemble)**.** *Suppose $\mathcal{X}$ is a set of size $|\mathcal{X}|$ with elements x, and suppose X is a random variable with probability density function $p_X(x)$. Suppose we have an ensemble $\{p_X(x), \sigma_x\}_{x\in\mathcal{X}}$ of quantum states where we encode each realization x into a quantum state σ_x. The expected density operator of the ensemble is*

$$\sigma \equiv \sum_{x\in\mathcal{X}} p_X(x)\sigma_x. \tag{15.2}$$

How can Alice transmit classical information reliably to Bob by making use of this ensemble? As suggested in the example from the previous section, Alice can select a subset of messages from the set $\mathcal{X}$, and Bob's task is to distinguish this subset of states as best he can. We equip him with certain tools: a *code* subspace projector Π and a set of *codeword* subspace projectors $\{\Pi_x\}_{x\in\mathcal{X}}$ with certain desirable properties (we explain these terms in more detail below). As a rough description, he can use these projectors to construct a quantum measurement that determines the message Alice sends. He would like to be almost certain that the received state lies in the subspace onto which the code subspace projector Π projects. He would also like to use the codeword subspace projectors $\{\Pi_x\}_{x\in\mathcal{X}}$ to determine the classical message that Alice sends. If the ensemble and the projectors satisfy certain conditions, the four conditions of the Packing Lemma, then it is possible for Bob to build up a measurement such that Alice can communicate reliably with him.

Suppose that Alice chooses some subset $\mathcal{C}$ of $\mathcal{X}$ for encoding classical information. The subset $\mathcal{C}$ that Alice chooses constitutes a *code*. Let us index the code $\mathcal{C}$ by a message set $\mathcal{M}$ with elements m. The set $\mathcal{M}$ contains messages m that Alice would like to transmit to Bob, and we assume that she chooses each message m with equal probability. The subensemble that Alice uses for transmitting classical information is thus as follows:

$$\left\{\frac{1}{|\mathcal{M}|}, \sigma_{c_m}\right\}, \tag{15.3}$$

where each c_m is a *codeword* that depends on the message m and takes a value in $\mathcal{X}$.

Bob needs a way to determine the classical message that Alice transmits. The most general way that quantum mechanics offers for retrieving classical information is a POVM. Thus, Bob performs some measurement described by a POVM $\{\Lambda_m\}_{m\in\mathcal{M}}$. Bob constructs this POVM by using the codeword subspace projectors $\{\Pi_x\}_{x\in\mathcal{X}}$ and the code subspace projector Π (we give the explicit construction in the proof of the Packing Lemma). If Alice transmits a message m, the probability that Bob correctly retrieves the message m is as follows:

$$\text{Tr}\{\Lambda_m \sigma_{c_m}\}. \tag{15.4}$$

Thus, the probability of error for a given message m while using the code $\mathcal{C}$ is as follows:

$$p_e(m, \mathcal{C}) \equiv 1 - \text{Tr}\{\Lambda_m \sigma_{c_m}\} \tag{15.5}$$

$$= \text{Tr}\{(I - \Lambda_m)\sigma_{c_m}\}. \tag{15.6}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We are interested in the performance of the code $\mathcal{C}$ that Alice and Bob choose, and we consider three different measures of performance.

1. The first and strongest measure of performance is the *maximal probability of error of the code* $\mathcal{C}$. A code $\mathcal{C}$ has maximum probability of error ϵ if the following criterion holds
$$\epsilon = \max_m p_e(m, \mathcal{C}). \tag{15.7}$$

2. A weaker measure of performance is the *average probability of error* $\bar{p}_e(\mathcal{C})$ *of the code* $\mathcal{C}$ where
$$\bar{p}_e(\mathcal{C}) \equiv \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} p_e(m, \mathcal{C}). \tag{15.8}$$

3. The third measure of performance is even weaker than the previous two but turns out to be the most useful in the mathematical proofs. It uses a conceptually different notion of code called a *random code*. Suppose that Alice and Bob choose a code $\mathcal{C}$ randomly from the set of all possible codes according to some probability density $p_{\mathcal{C}}$ (the code $\mathcal{C}$ itself therefore becomes a random variable!) The third measure of performance is the *expectation of the average probability of error of a random code* $\mathcal{C}$ where the expectation is with respect to the set of all possible codes with message set $\mathcal{M}$ chosen according to the density $p_{\mathcal{C}}$:
$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\} \equiv \mathbb{E}_{\mathcal{C}}\left\{ \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} p_e(m, \mathcal{C}) \right\} \tag{15.9}$$
$$= \sum_{\mathcal{C}} p_{\mathcal{C}} \left(\frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} p_e(m, \mathcal{C}) \right). \tag{15.10}$$
We will see that considering this performance criterion simplifies the mathematics in the proof of the Packing Lemma. Then we will employ a series of arguments to strengthen the result for this weakest performance criterion to the first and strongest performance criterion.

15.3 Statement of the Packing Lemma

Lemma 15.3.1 (Packing Lemma)**.** *Suppose that we have an ensemble as in Definition 15.2.1. Suppose that a code subspace projector* Π *and codeword subspace projectors* $\{\Pi_x\}_{x\in\mathcal{X}}$ *exist, they project onto subspaces of* $\mathcal{H}$*, and these projectors and the ensemble satisfy the following*

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

conditions:

$$\text{Tr}\{\Pi\sigma_x\} \geq 1-\epsilon, \tag{15.11}$$
$$\text{Tr}\{\Pi_x\sigma_x\} \geq 1-\epsilon, \tag{15.12}$$
$$\text{Tr}\{\Pi_x\} \leq d, \tag{15.13}$$
$$\Pi\sigma\Pi \leq \frac{1}{D}\Pi, \tag{15.14}$$

where $0 < d < D$. Suppose that $\mathcal{M}$ is a set of size $|\mathcal{M}|$ with elements m. We generate a set $\mathcal{C} = \{C_m\}_{m\in\mathcal{M}}$ of random variables C_m where each random variable C_m corresponds to the message m, has density $p_X(x)$ so that its distribution is independent of the particular message m, and takes a value in $\mathcal{X}$. This set constitutes a random code. Then there exists a corresponding POVM $(\Lambda_m)_{m\in\mathcal{M}}$ that reliably distinguishes between the states $(\sigma_{C_m})_{m\in\mathcal{M}}$ in the sense that the expectation of the average probability of detecting the correct state is high:

$$\mathbb{E}_{\mathcal{C}}\left\{\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\text{Tr}\{\Lambda_m\sigma_{C_m}\}\right\} \geq 1-2\left(\epsilon+2\sqrt{\epsilon}\right)-4\left(\frac{D}{d|\mathcal{M}|}\right)^{-1}, \tag{15.15}$$

given that D/d is large, $|\mathcal{M}| \ll D/d$, and ϵ is arbitrarily small.

Condition (15.11) states that the code subspace with projector Π contains each message σ_x with high probability. Condition (15.12) states that each codeword subspace projector Π_x contains its corresponding state σ_x with high probability. Condition (15.13) states that the dimension of each codeword subspace projector Π_x is less than some positive number d. Condition (15.14) states that the distribution of the ensemble with expected density operator σ is approximately uniform when projecting it onto the subspace with projector Π. Conditions (15.11) and (15.14) imply that

$$\text{Tr}\{\Pi\} \geq D(1-\epsilon), \tag{15.16}$$

so that the dimension of the code subspace projector Π is approximately D. We show how to construct a code with messages that Alice wants to send. These four conditions are crucial for constructing a decoding POVM with the desirable property that it can distinguish between the messages with high probability.

The main idea of the Packing Lemma is that we can pack $|\mathcal{M}|$ classical messages into a subspace with corresponding projector Π. There is then a small probability of error when trying to detect the classical messages with codeword subspace projectors Π_x. The intuition is the same as that depicted in Figure 2.6. We are trying to pack as many subspaces of size d into a larger space of size D. In the proof of the HSW coding theorem in Chapter 19, D will be of size $\approx 2^{nH(B)}$ and d will be of size $\approx 2^{nH(B|X)}$, suggesting that we pack in $\approx 2^{n[H(B)-H(B|X)]} = 2^{nI(X;B)}$ messages while still being able to distinguish them reliably.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

15.4 Proof of the Packing Lemma

The proof technique employs a Shannon-like argument where we generate a code at random. We show how to construct a POVM, the "pretty-good" measurement, that can decode a classical message with high probability. We then prove that the expectation of the average error probability is small (where the expectation is over all random codes). In a corollary in the next section, we finally use standard Shannon-like arguments to show that a code exists whose maximal probability of error for all messages is small.

15.4.1 Code Construction

We present a Shannon-like random coding argument to simplify the mathematics that follow. We construct a code $\mathcal{C}$ at random by independently generating $|\mathcal{M}|$ codewords according to the distribution $p_X(x)$. Let $\mathcal{C} \equiv \{c_m\}_{m\in\mathcal{M}}$ be a collection of the realizations c_m of $|\mathcal{M}|$ independent random variables C_m. Each C_m takes a value c_m in $\mathcal{X}$ with probability $p_X(c_m)$ and represents a classical codeword in the random code $\mathcal{C}$. The probability $p(\mathcal{C})$ of choosing a particular code $\mathcal{C}$ is equal to the following:

$$p(\mathcal{C}) = \prod_{m=1}^{|\mathcal{M}|} p_X(c_m). \tag{15.17}$$

There is a great advantage to choosing the code in this way. The expectation of any product $f(C_m)g(C_{m'})$ of two functions f and g of two different random codewords C_m and $C_{m'}$, where the expectation is with respect to the random choice of code, factors as follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}\{f(C_m)g(C_{m'})\} &= \sum_{c} p(c) f(c_m) g(c_{m'}) && (15.18)\\
&= \sum_{c_1\in\mathcal{X}} p_X(c_1) \cdots \sum_{c_{|\mathcal{M}|}\in\mathcal{X}} p_X\left(c_{|\mathcal{M}|}\right) f(c_m) g(c_{m'}) && (15.19)\\
&= \sum_{c_m\in\mathcal{X}} p_X(c_m) f(c_m) \sum_{c_{m'}\in\mathcal{X}} p_X(c_{m'}) g(c_{m'}) && (15.20)\\
&= \mathbb{E}_X\{f(X)\}\mathbb{E}_X\{g(X)\}. && (15.21)
\end{aligned}$$

This factoring happens because of the random way in which we choose the code, and we exploit this fact in the proof of the Packing Lemma. We employ the following events in sequence:

1. We choose a random code as described above.
2. We reveal the code to the sender and receiver.
3. The sender chooses a message m at random (with uniform probability according to some random variable M) from $\mathcal{M}$ and encodes it in the codeword c_m. The quantum state that the sender transmits is then equal to σ_{c_m}.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

4. The receiver performs the POVM $(\Lambda_m)_{m\in\mathcal{M}}$ to determine the message that the sender transmits, and each POVM element Λ_m corresponds to a message m in the code. The receiver obtains a classical result from the measurement, and we model it with the random variable M'. The conditional probability $\Pr\{M' = m \mid M = m\}$ of obtaining the correct result from the measurement is equal to

$$\Pr\{M' = m \mid M = m\} = \text{Tr}\{\Lambda_m \sigma_{c_m}\}. \tag{15.22}$$

5. The receiver decodes correctly if $M' = M$ and decodes incorrectly if $M' \neq M$.

15.4.2 POVM Construction

We cannot directly use the projectors Π_x in a POVM because they do not satisfy the conditions for being a POVM. Namely, it is not necessarily true that $\sum_{x\in\mathcal{X}} \Pi_x = I$. Also, the codeword subspace projectors Π_x may have support outside that of the code subspace projector Π.

To remedy these problems, first consider the following set of operators:

$$\forall x \quad \Upsilon_x \equiv \Pi\Pi_x\Pi. \tag{15.23}$$

The operator Υ_x is a positive operator, and the effect of "coating" the codeword subspace projector Π_x with the code subspace projector Π is to slice out any part of the support Π_x that is not in the support of Π. From the conditions (15.11-15.12) of the Packing Lemma, there should be little probability for our states of interest to lie in the part of the support of Π_x outside the support of Π. The operators Υ_x have the desirable property that they only have support inside of the subspace corresponding to the code subspace projector Π. So we have remedied the second problem stated above.

We now remedy the first problem stated above by constructing a POVM $\{\Lambda_m\}_{m\in\mathcal{M}}$ with the following elements:

$$\Lambda_m \equiv \left(\sum_{m'=1}^{|\mathcal{M}|} \Upsilon_{c_{m'}}\right)^{-\frac{1}{2}} \Upsilon_{c_m} \left(\sum_{m'=1}^{|\mathcal{M}|} \Upsilon_{c_{m'}}\right)^{-\frac{1}{2}}. \tag{15.24}$$

The above POVM is the "pretty-good" or "square-root" measurement. The POVM elements also have the property that $\sum_{m=1}^{|\mathcal{M}|} \Lambda_m \leq I$. Note that the inverse square root $A^{-\frac{1}{2}}$ of an operator A is defined as the inverse square root operation only on the support of A. That is, given a spectral decomposition of the operator A so that

$$A = \sum_a a|a\rangle\langle a|, \tag{15.25}$$

and

$$A^{-\frac{1}{2}} = \sum_a f(a)|a\rangle\langle a|, \tag{15.26}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where

$$f(a) = \begin{cases} a^{-\frac{1}{2}} & : \ a \neq 0 \\ 0 & : \ a = 0 \end{cases}. \tag{15.27}$$

We can have a complete POVM by adding the operator $\Lambda_0 = I - \sum_m \Lambda_m$ to the set. The idea of the pretty-good measurement is that the POVM elements $\{\Lambda_m\}_{m=1}^{|\mathcal{M}|}$ correspond to the messages sent and the element Λ_0 corresponds to an error result.

The above square root measurement is useful because we can apply the following result that holds for any positive operators S and T such that $0 \leq S \leq I$ and $T \geq 0$:

$$I - (S+T)^{-\frac{1}{2}} S (S+T)^{-\frac{1}{2}} \leq 2(I-S) + 4T. \tag{15.28}$$

Exercise 15.4.1 (Hayashi-Nagaoka Operator Inequality) Prove the following inequality

$$I - (S+T)^{-\frac{1}{2}} S (S+T)^{-\frac{1}{2}} \leq 2(I-S) + 4T \tag{15.29}$$

that holds for any positive operators S and T such that $0 \leq S \leq I$ and $T \geq 0$. (Hint: Suppose that the projection onto the range of $S+T$ is the whole of Hilbert space. Use the fact that

$$(M - 2I)T(M - 2I) \geq 0 \tag{15.30}$$

for a positive operator T and any operator M. Use the fact that $\sqrt{\cdot}$ is operator monotone: $A \geq B \Rightarrow \sqrt{A} \geq \sqrt{B}$.)

We make the following substitutions:

$$T = \sum_{m' \neq m}^{|\mathcal{M}|} \Upsilon_{c_{m'}}, \qquad S = \Upsilon_{c_m}, \tag{15.31}$$

so that the bound in (15.28) becomes

$$I - \Lambda_m \leq 2(I - \Upsilon_{c_m}) + 4 \sum_{m' \neq m}^{|\mathcal{M}|} \Upsilon_{c_{m'}}. \tag{15.32}$$

The above expression is useful in our error analysis in the next section.

15.4.3 Error Analysis

Suppose we have chosen a particular code $\mathcal{C}$. Let $p_e(m, \mathcal{C})$ be the probability of decoding incorrectly given that message m was sent while using the code $\mathcal{C}$:

$$p_e(m, \mathcal{C}) \equiv \text{Tr}\{(I - \Lambda_m)\sigma_{c_m}\}. \tag{15.33}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Then using (15.28), the following bound applies to the average error probability:

$$p_e(m,\mathcal{C}) \leq 2\ \mathrm{Tr}\{(I-\Upsilon_{c_m})\sigma_{c_m}\} + 4\ \mathrm{Tr}\left\{\sum_{m'\neq m}^{|\mathcal{M}|}\Upsilon_{c_{m'}}\sigma_{c_m}\right\} \tag{15.34}$$

$$= 2\ \mathrm{Tr}\{(I-\Upsilon_{c_m})\sigma_{c_m}\} + 4\sum_{m'\neq m}^{|\mathcal{M}|}\mathrm{Tr}\{\Upsilon_{c_{m'}}\sigma_{c_m}\} \tag{15.35}$$

The above bound on the message error probability for code $\mathcal{C}$ has a similar interpretation as that in classical Shannon-like proofs. We bound the error probability by the probability of decoding to any message outside the message space operator Υ_{c_m} (the first term in (15.35)) summed with the probability of confusing the transmitted message with a message $c_{m'}$ different from the correct one (the second term in (15.35)). The average error probability $\bar{p}_e(\mathcal{C})$ over all transmitted messages for code $\mathcal{C}$ is

$$\bar{p}_e(\mathcal{C}) = \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|} p_e(m,\mathcal{C}), \tag{15.36}$$

because Alice chooses the message m that she would like to transmit according to the uniform distribution. The average error probability $\bar{p}_e(\mathcal{C})$ then obeys the following bound:

$$\bar{p}_e(\mathcal{C}) \leq \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\left[2\ \mathrm{Tr}\{(I-\Upsilon_{c_m})\sigma_{c_m}\} + 4\sum_{m'\neq m}^{|\mathcal{M}|}\mathrm{Tr}\{\Upsilon_{c_{m'}}\sigma_{c_m}\}\right]. \tag{15.37}$$

Consider the first term $\mathrm{Tr}\{(I-\Upsilon_{c_m})\sigma_{c_m}\}$ on the RHS above. We can bound it from above by a small number, simply by applying (15.11-15.12) and the Gentle Operator Lemma (Lemma 9.4.2). Consider the following chain of inequalities:

$$\mathrm{Tr}\{\Upsilon_{c_m}\sigma_{c_m}\} = \mathrm{Tr}\{\Pi\Pi_{c_m}\Pi\sigma_{c_m}\} \tag{15.38}$$

$$= \mathrm{Tr}\{\Pi_{c_m}\Pi\sigma_{c_m}\Pi\} \tag{15.39}$$

$$\geq \mathrm{Tr}\{\Pi_{c_m}\sigma_{c_m}\} - \|\Pi\sigma_{c_m}\Pi - \sigma_{c_m}\|_1 \tag{15.40}$$

$$\geq 1-\epsilon-2\sqrt{\epsilon} \tag{15.41}$$

The first equality follows by the definition of Υ_{c_m} in (15.23). The second equality follows from cyclicity of the trace. The first inequality follows from applying (9.58). The last inequality follows from applying (15.11) to $\mathrm{Tr}\{\Pi_{c_m}\sigma_{c_m}\}$ and applying (15.12) and the Gentle Operator Lemma (Lemma 9.4.2) to $\|\Pi\sigma_{c_m}\Pi-\sigma_{c_m}\|_1$. The above bound then implies the following one:

$$\mathrm{Tr}\{(I-\Upsilon_{c_m})\sigma_{c_m}\} = 1-\mathrm{Tr}\{\Upsilon_{c_m}\sigma_{c_m}\} \tag{15.42}$$

$$\leq \epsilon+2\sqrt{\epsilon}, \tag{15.43}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and by substituting into (15.37), we get the following bound on the average probability of error:

$$\bar{p}_e(\mathcal{C}) \leq \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\left[2\left(\epsilon + 2\sqrt{\epsilon}\right) + 4\sum_{m'\neq m}^{|\mathcal{M}|}\operatorname{Tr}\{\Upsilon_{c_{m'}}\sigma_{c_m}\}\right] \tag{15.44}$$

$$= 2\left(\epsilon + 2\sqrt{\epsilon}\right) + \frac{4}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\sum_{m'\neq m}^{|\mathcal{M}|}\operatorname{Tr}\{\Upsilon_{c_{m'}}\sigma_{c_m}\}. \tag{15.45}$$

At this point, bounding the average error probability further is a bit difficult, given the sheer number of combinations of terms $\operatorname{Tr}\{\Upsilon_{c_{m'}}\sigma_{c_m}\}$ that we would have to consider to do so. Thus, we should now invoke the classic Shannon argument in order to simplify the mathematics. Instead of considering the average probability of error, we consider the expectation of the average error probability $\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\}$ with respect to all possible random codes $\mathcal{C}$. Considering this error quantity significantly simplifies the mathematics because of the way in which we constructed the code. We can use the probability distribution $p_X(x)$ to compute the expectation $\mathbb{E}_{\mathcal{C}}$ because we constructed our code according to this distribution. The bound above becomes as follows:

$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\} \leq \mathbb{E}_{\mathcal{C}}\left\{2\left(\epsilon + 2\sqrt{\epsilon}\right) + \frac{4}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\sum_{m'\neq m}^{|\mathcal{M}|}\operatorname{Tr}\{\Upsilon_{c_{m'}}\sigma_{c_m}\}\right\} \tag{15.46}$$

$$= 2\left(\epsilon + 2\sqrt{\epsilon}\right) + \frac{4}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\sum_{m'\neq m}^{|\mathcal{M}|}\mathbb{E}_{\mathcal{C}}\{\operatorname{Tr}\{\Upsilon_{c_{m'}}\sigma_{c_m}\}\}, \tag{15.47}$$

by exploiting the linearity of expectation.

We now calculate the expectation of the expression $\operatorname{Tr}\{\Upsilon_{C_{m'}}\sigma_{C_m}\}$ over all random codes $\mathcal{C}$:

$$\mathbb{E}_{\mathcal{C}}\{\operatorname{Tr}\{\Upsilon_{C_{m'}}\sigma_{C_m}\}\} = \mathbb{E}_{\mathcal{C}}\{\operatorname{Tr}\{\Pi\Pi_{C_{m'}}\Pi\sigma_{C_m}\}\} \tag{15.48}$$

$$= \mathbb{E}_{\mathcal{C}}\{\operatorname{Tr}\{\Pi_{C_{m'}}\Pi\sigma_{C_m}\Pi\}\}. \tag{15.49}$$

The first equality follows from the definition in (15.23), and the second equality follows from cyclicity of trace. Independence of random variables C_m and $C_{m'}$ (from the code construction) gives that the above expression equals

$$\operatorname{Tr}\{\mathbb{E}_{\mathcal{C}}\{\Pi_{C_{m'}}\}\Pi\mathbb{E}_{\mathcal{C}}\{\sigma_{C_m}\}\Pi\} = \operatorname{Tr}\{\mathbb{E}_{\mathcal{C}}\{\Pi_{C_{m'}}\}\Pi\sigma\Pi\} \tag{15.50}$$

$$\leq \operatorname{Tr}\left\{\mathbb{E}_{\mathcal{C}}\{\Pi_{C_{m'}}\}\frac{1}{D}\Pi\right\} \tag{15.51}$$

$$= \frac{1}{D}\operatorname{Tr}\{\mathbb{E}_{\mathcal{C}}\{\Pi_{C_{m'}}\}\Pi\}. \tag{15.52}$$

where the first equality uses the fact that $\mathbb{E}_{\mathcal{C}}\{\sigma_{C_m}\} = \sum_{x\in\mathcal{X}} p(x)\sigma_x = \sigma$ and Π is a constant with respect to the expectation. The first inequality uses the fourth condition (15.14) of the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Packing Lemma, the fact that $\Pi\sigma\Pi$, Π, and $\Pi_{C_{m'}}$ are all positive operators, and $\text{Tr}\{CA\} \geq \text{Tr}\{CB\}$ for $C \geq 0$ and $A \geq B$. Continuing, we have

$$\frac{1}{D}\text{Tr}\{\mathbb{E}_{\mathcal{C}}\{\Pi_{C_{m'}}\}\Pi\} \leq \frac{1}{D}\text{Tr}\{\mathbb{E}_{\mathcal{C}}\{\Pi_{C_{m'}}\}\} \tag{15.53}$$

$$= \frac{1}{D}\mathbb{E}_{\mathcal{C}}\{\text{Tr}\{\Pi_{C_{m'}}\}\} \tag{15.54}$$

$$\leq \frac{d}{D}. \tag{15.55}$$

The first inequality follows from the fact that $\Pi \leq I$ and $\Pi_{C_{m'}}$ is a positive operator. The last inequality follows from (15.13). The following inequality then holds by considering the development from (15.48) to (15.55):

$$\mathbb{E}_{\mathcal{C}}\{\text{Tr}\{\sigma_{C_m}\Upsilon_{C_{m'}}\}\} \leq \frac{d}{D}. \tag{15.56}$$

We substitute into (15.35) to show that the expectation $\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\}$ of the average error probability $\bar{p}_e(\mathcal{C})$ over all codes obeys

$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\} \leq 2\left(\epsilon + 2\sqrt{\epsilon}\right) + \frac{4}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\sum_{m'\neq m}^{|\mathcal{M}|}\mathbb{E}_{\mathcal{C}}\{\text{Tr}\{\sigma_{c_m}\Upsilon_{c_{m'}}\}\} \tag{15.57}$$

$$\leq 2\left(\epsilon + 2\sqrt{\epsilon}\right) + \frac{4}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\sum_{m'\neq m}^{|\mathcal{M}|}\frac{d}{D} \tag{15.58}$$

$$\leq 2\left(\epsilon + 2\sqrt{\epsilon}\right) + 4(|\mathcal{M}| - 1)\frac{d}{D} \tag{15.59}$$

$$\leq 2\left(\epsilon + 2\sqrt{\epsilon}\right) + 4\left(\frac{D}{d|\mathcal{M}|}\right)^{-1}. \tag{15.60}$$

15.5 Derandomization and Expurgation

The above version of the Packing Lemma is a randomized version that shows how the expectation of the average probability of error is small. We now prove a derandomized version that guarantees the existence of a code with small maximal error probability for each message. The last two arguments are traditionally called *derandomization* and *expurgation*.

Corollary 15.5.1. *Suppose we have the ensemble as in Definition 15.2.1. Suppose that a code subspace projector Π and codeword subspace projectors $\{\Pi_x\}_{x\in\mathcal{X}}$ exist, they project onto subspaces of $\mathcal{H}$, and these projectors and the ensemble have the following properties:*

$$\text{Tr}\{\Pi\sigma_x\} \geq 1 - \epsilon, \tag{15.61}$$

$$\text{Tr}\{\Pi_x\sigma_x\} \geq 1 - \epsilon, \tag{15.62}$$

$$\text{Tr}\{\Pi_x\} \leq d, \tag{15.63}$$

$$\Pi\sigma\Pi \leq \frac{1}{D}\Pi, \tag{15.64}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $0 < d < D$. Suppose that $\mathcal{M}$ is a set of size $|\mathcal{M}|$ with elements m. Then there exists a code $\mathcal{C}_0 = \{c_m\}_{m\in\mathcal{M}}$ with codewords c_m depending on the message m and taking values in $\mathcal{X}$ and there exists a corresponding POVM $(\Lambda_m)_{m\in\mathcal{M}}$ that reliably distinguishes between the states $(\sigma_{c_m})_{m\in\mathcal{M}}$ in the sense that the probability of detecting the correct state is high:

$$\forall m \in \mathcal{M} \qquad \operatorname{Tr}\{\Lambda_m \sigma_{c_m}\} \geq 1 - 4\left(\epsilon + 2\sqrt{\epsilon}\right) - 8\left(\frac{D}{d|\mathcal{M}|}\right)^{-1}, \tag{15.65}$$

because we can make ϵ and $\left(\frac{D}{d|\mathcal{M}|}\right)^{-1}$ arbitrarily small (this holds if $|\mathcal{M}| \ll D/d$). We can use the code $\mathcal{C}_0$ and the POVM $(\Lambda_m)_{m\in\mathcal{M}}$ respectively to encode and decode $|\mathcal{M}|$ classical messages with high success probability.

Proof. Generate a random code according to the construction in the previous lemma. The expectation of the average error probability then satisfies the bound in the Packing Lemma. We now make a few standard Shannon-like arguments to strengthen the result of the previous lemma. **Derandomization.** The expectation of the average error probability $\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\}$ satisfies the following bound:

$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\} \leq \epsilon'. \tag{15.66}$$

It then follows that the average error probability of at least one code $\mathcal{C}_0 = \{c_m\}_{m\in\mathcal{M}}$ satisfies the above bound:

$$\bar{p}_e(\mathcal{C}_0) \leq \epsilon'. \tag{15.67}$$

Choose this code $\mathcal{C}_0$ as the code, and it is possible to find this code $\mathcal{C}_0$ in practice by exhaustive search. This process is known as *derandomization*.

Exercise 15.5.1 Use Markov's inequality to prove an even stronger derandomization of the Packing Lemma. Prove that the overwhelming fraction $1-\sqrt{\epsilon'}$ of codes contructed randomly have average error probability less than $\sqrt{\epsilon'}$.

Expurgation. We now consider the maximal error probability instead of the average error probability by an expurgation argument. We know that $p_e(m) \leq 2\epsilon'$ for at least half of the indices (if it were not true, then these indices would contribute more than ϵ' to the average error probability $\bar{p}_e$). Throw out the half of the codewords with the worst decoding probability and redefine the code according to the new set of indices. These steps have a negligible effect on the parameters of the code when we later consider a large number of uses of a noisy quantum channel. It is helpful to refer back to Exercise 2.2.1 at this point. □

Exercise 15.5.2 Use Markov's inequality to prove an even stronger expurgation argument (following on the result of Exercise 15.5.1). Prove that we can retain a large fraction $1-\sqrt[4]{\epsilon'}$ of the codewords (expurgating $\sqrt[4]{\epsilon'}$ of them) so that each remaining codeword has error probability less than $\sqrt[4]{\epsilon'}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 15.5.3 Prove that the Packing Lemma and its corollary hold for the same ensemble and a set of projectors for which the following conditions hold:

$$\sum_{x\in\mathcal{X}} p_X(x)\text{Tr}\{\sigma_x\Pi\} \geq 1-\epsilon, \tag{15.68}$$

$$\sum_{x\in\mathcal{X}} p_X(x)\text{Tr}\{\sigma_x\Pi_x\} \geq 1-\epsilon, \tag{15.69}$$

$$\text{Tr}\{\Pi_x\} \leq d, \tag{15.70}$$

$$\Pi\sigma\Pi \leq \frac{1}{D}\Pi, \tag{15.71}$$

Exercise 15.5.4 Prove that a variation of the Packing Lemma holds in which the POVM is of the following form:

$$\Lambda_m \equiv \left(\sum_{m'=1}^{|\mathcal{M}|}\Pi_{c_{m'}}\right)^{-\frac{1}{2}}\Pi_{c_m}\left(\sum_{m'=1}^{|\mathcal{M}|}\Pi_{c_{m'}}\right)^{-\frac{1}{2}}. \tag{15.72}$$

That is, it is not actually necessary to "coat" each operator in the square-root measurement with the overall message subspace projector.

15.6 History and Further Reading

Holevo [144], Schumacher, and Westmoreland [219] did not prove the classical coding theorem with the Packing Lemma, but they instead used other arguments to bound the probability of error. The operator inequality in (15.28) is at the heart of the Packing Lemma. Hayashi and Nagaoka proved this operator inequality in the more general setting of the quantum information spectrum method [130], where there is no IID constraint and essentially no structure to a channel. Devetak *et al.* later exploited this operator inequality in the context of entanglement-assisted classical coding [156] and followed the approach in Ref. [130] to prove the Packing Lemma.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 16

The Covering Lemma

The goal of the Covering Lemma is perhaps opposite to that of the Packing Lemma because it applies in a setting where one party wishes to make messages *indistinguishable* to another party (instead of trying to make them distinguishable as in the Packing Lemma of the previous chapter). That is, the Covering Lemma is helpful when one party is trying to simulate a noisy channel to another party, rather than trying to simulate a noiseless channel. One party can accomplish this task by randomly covering the Hilbert space of the other party (this viewpoint gives the Covering Lemma its name).

One can certainly simulate noise by choosing a quantum state uniformly at random from a large set of quantum states and passing along the chosen quantum state to a third party without telling which state was chosen. But the problem with this approach is that it could potentially be expensive if the set from which we choose a random state is large, and we would really like to use as few resources as possible in order to simulate noise. That is, we would like the set from which we choose a quantum state uniformly at random to be as small as possible when simulating noise. The Covering Lemma is similar to the Packing Lemma in the sense that its conditions for application are general (involving bounds on projectors and an ensemble), but it gives an asymptotically efficient scheme for simulating noise when we apply it in an IID setting.

One application of the Covering Lemma in quantum Shannon theory is in the construction of a code for transmitting private classical information over a quantum channel (discussed in Chapter 22). The method of proof for private classical transmission involves a clever combination of packing messages so that Bob can distinguish them, while covering Eve's space in such a way that Eve cannot distinguish the messages intended for Bob. A few other applications of the Covering Lemma are in secret key distillation, determining the amount of noise needed to destroy correlations in a bipartite state, and compressing the outcomes of an IID measurement on an IID quantum state.

We begin this chapter with a simple example to explain the main idea behind the Covering Lemma. Section 16.2 then discusses its general setting and gives its statement. We dissect its proof into several different parts: the construction of a "Chernoff ensemble," the construction of a "Chernoff code," the application of the Chernoff bound, and the error analysis. The

main tool that we use to prove the Covering Lemma is the Operator Chernoff bound. This bound is a generalization of the standard Chernoff bound from probability theory, which states that the sample mean of a sequence of IID random variables converges exponentially to its true mean. The proof of the operator version of the Chernoff bound is straightforward and we provide it in Appendix A. The exponential convergence rate in the Chernoff bound is much stronger than the polynomial convergence rate from Chebyshev's inequality and is helpful in proving the existence of good private classical codes in Chapter 22.

16.1 Introductory Example

Suppose that Alice is trying to communicate with Bob as before, but now there is an eavesdropper Eve listening in on their communication. Alice wants the messages that she is sending to Bob to be *private* so that Eve does not gain any information about the message that she is sending.

How can Alice make the information that she is sending private? The strongest criterion for security is to ensure that whatever Eve receives is independent of what Alice is sending. Alice may have to sacrifice the amount of information she can communicate to Bob in order to have privacy, but this sacrifice is worth it to her because she really does not want Eve to know anything about the intended message for Bob.

We first give an example to motivate a general method that Alice can use to make her information private. Suppose Alice can transmit one of four messages $\{a, b, c, d\}$ to Bob, and suppose he receives them perfectly as distinguishable quantum states. She chooses from these messages with equal probability. Suppose further that Alice and Eve know that Eve receives one of the following four states corresponding to each of Alice's messages:

$$a \rightarrow |0\rangle, \qquad b \rightarrow |1\rangle, \qquad c \rightarrow |+\rangle, \qquad d \rightarrow |-\rangle. \tag{16.1}$$

Observe that each of Eve's states lies in the two-dimensional Hilbert space of a qubit. We refer to the quantum states in the above ensemble as "Eve's ensemble."

We are not so much concerned for what Bob receives for the purposes of this example, but we just make the assumption that he can distinguish the four messages that Alice sends. Without loss of generality, let us just assume that he receives the messages unaltered in some preferred orthonormal basis such as $\{|a\rangle, |b\rangle, |c\rangle, |d\rangle\}$ so that he can distinguish the four messages, and let us call this ensemble "Bob's ensemble."

Both Alice and Eve then know that the expected density operator of Eve's ensemble is the maximally mixed state if Eve does not know which message Alice chooses:

$$\frac{1}{4}|0\rangle\langle 0| + \frac{1}{4}|1\rangle\langle 1| + \frac{1}{4}|+\rangle\langle +| + \frac{1}{4}|-\rangle\langle -| = \frac{I}{2}. \tag{16.2}$$

How can Alice ensure that Eve's information is independent of the message Alice is sending? Alice can choose subsets or subensembles of the states in Eve's ensemble to simulate the expected density operator of Eve's ensemble. Let us call these new simulating ensembles the "fake ensembles." Alice chooses the member states of the fake ensembles according to the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

uniform distribution in order to randomize Eve's knowledge. The density operator for each new fake ensemble is its "fake expected density operator."

Which states work well for being members of the fake ensembles? An equiprobable mixture of the states $|0\rangle$ and $|1\rangle$ suffices to simulate the expected density operator of Eve's ensemble because the fake expected density operator of this new ensemble is as follows:

$$\frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}|1\rangle\langle 1| = \frac{I}{2}. \tag{16.3}$$

An equiprobable mixture of the states $|+\rangle$ and $|-\rangle$ also works because the fake expected density operator of this other fake ensemble is as follows:

$$\frac{1}{2}|+\rangle\langle +| + \frac{1}{2}|-\rangle\langle -| = \frac{I}{2}. \tag{16.4}$$

So it is possible for Alice to encode a private bit this way. She first generates a random bit that selects a particular message within each fake ensemble. So she selects a or b according to the random bit if she wants to transmit a "0" privately to Bob, and she selects c or d according to the random bit if she wants to transmit a "1" privately to Bob. In each of these cases, Eve's resulting expected density operator is the maximally mixed state. Thus, there is no measurement that Eve can perform to distinguish the original message that Alice transmits. Bob, on the other hand, can perform a measurement in the basis $\{|a\rangle, |b\rangle, |c\rangle, |d\rangle\}$ to determine Alice's private bit. Then Eve's best strategy is just to guess at the transmitted message. In the case of one private bit, Eve can guess its value correctly with probability $1/2$, but Alice and Bob can make this probability exponentially small if Alice sends more private bits with this technique (the guessing probability becomes $\frac{1}{2^n}$ for n private bits).

We can explicitly calculate Eve's accessible information about the private bit. Consider Eve's impression of the state if she does not know which message Alice transmits—it is an equal mixture of the following states: $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$ (the maximally mixed state $I/2$). Eve's impression of the state "improves" to an equal mixture of the states $\{|0\rangle, |1\rangle\}$ or $\{|+\rangle, |-\rangle\}$, both with density operator $I/2$, if she does know which message Alice transmits. The following classical-quantum state describes this setting:

$$\rho^{KME} \equiv \frac{1}{4}\left[\begin{array}{c} |0\rangle\langle 0|^K \otimes |0\rangle\langle 0|^M \otimes |0\rangle\langle 0|^E + |0\rangle\langle 0|^K \otimes |1\rangle\langle 1|^M \otimes |1\rangle\langle 1|^E + \\ |1\rangle\langle 1|^K \otimes |0\rangle\langle 0|^M \otimes |+\rangle\langle +|^E + |1\rangle\langle 1|^K \otimes |1\rangle\langle 1|^M \otimes |-\rangle\langle -|^E \end{array} \right], \tag{16.5}$$

where we suppose that Eve never has access to the M register. If she does not have access to K, then her state is the maximally mixed state (obtained by tracing out K and M). If she does know K, then her state is still the maximally mixed state. We can now calculate Eve's accessible information about the private bit by evaluating the quantum mutual information

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

of the state ρ^{KME}:

$$I(K;E)_\rho = H(E)_\rho - H(E|K)_\rho \tag{16.6}$$

$$= H\left(\frac{I}{2}\right) - \sum_{k=0}^{1} \frac{1}{2} H(E)_{\rho_k} \tag{16.7}$$

$$= H\left(\frac{I}{2}\right) - \frac{1}{2} H(\{|0\rangle, |1\rangle\}) - \frac{1}{2} H(\{|+\rangle, |-\rangle\}) \tag{16.8}$$

$$= H\left(\frac{I}{2}\right) - \frac{1}{2} H\left(\frac{I}{2}\right) - \frac{1}{2} H\left(\frac{I}{2}\right) = 0. \tag{16.9}$$

Thus using this scheme, Eve has no accessible information about the private bit as we argued before.

We are interested in making this scheme use as little noise as possible because Alice would like to transmit as much information as she can to Bob while still retaining privacy. Therefore, Alice should try to make the fake ensembles as small as possible. In the above example, Alice cannot make the fake ensembles any smaller because a smaller size would leak information to Eve.

16.2 Setting and Statement of the Covering Lemma

The setting of the Covering Lemma is a generalization of the setting in the above example. It essentially uses the same strategy for making information private, but the mathematical analysis becomes more involved in the more general setting. In general, we cannot have perfect privacy as in the above example, but instead we ask only for approximate privacy. Approximate privacy then becomes perfect in the asymptotic limit in the IID setting.

We first define the relevant ensemble for the Covering Lemma. We call it the "true ensemble" in order to distinguish it from the "fake ensemble."

Definition 16.2.1 (True Ensemble)**.** *Suppose $\mathcal{X}$ is a set of size $|\mathcal{X}|$ with elements x. Suppose we have an ensemble $\{p_X(x), \sigma_x\}_{x\in\mathcal{X}}$ of quantum states where each value x occurs with probability $p_X(x)$ according to some random variable X, and suppose we encode each value x into a quantum state σ_x. The expected density operator of the ensemble is $\sigma \equiv \sum_{x\in\mathcal{X}} p_X(x)\sigma_x$.*

The definition for a fake ensemble is similar to the way that we constructed the fake ensembles in the example. It is merely a subset of the states in the true ensemble chosen according to a uniform distribution.

Definition 16.2.2 (Fake Ensemble)**.** *Consider a set $\mathcal{S}$ where $\mathcal{S} \subset \mathcal{X}$. The fake ensemble is as follows:*

$$\left\{\frac{1}{|\mathcal{S}|}, \sigma_s\right\}_{s\in\mathcal{S}}. \tag{16.10}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Let $\overline{\sigma}$ denote the "fake expected density operator" of the fake ensemble:

$$\overline{\sigma}(\mathcal{S}) \equiv \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \sigma_s. \tag{16.11}$$

In the example, Alice was able to obtain perfect privacy from Eve. We need a good measure of privacy because it is not possible in general to obtain perfect privacy, but Alice can instead obtain only approximate privacy. We call this measure the "obfuscation error" because it determines how well Alice can obfuscate the state that Eve receives.

Definition 16.2.3 (Obfuscation Error)**.** *The obfuscation error $o_e(\mathcal{S})$ of set $\mathcal{S}$ is a measure of how close the fake expected density operator $\overline{\sigma}(\mathcal{S})$ is to the actual expected density operator:*

$$o_e(\mathcal{S}) = \|\overline{\sigma}(\mathcal{S}) - \sigma\|_1. \tag{16.12}$$

The goal for Alice is to make the size of her fake ensembles as small as possible while still having privacy from Eve. The covering lemma makes this tradeoff exact by determining exactly how small each fake ensemble can be in order to obtain a certain obfuscation error.

The hypotheses of the Covering Lemma are somewhat similar to those of the Packing Lemma. But as stated in the introduction of this chapter, the goal of the Covering Lemma is much different.

Lemma 16.2.1 (Covering Lemma)**.** *Suppose we are given an ensemble as defined in Definition 16.2.1. Suppose a total subspace projector Π and codeword subspace projectors $\{\Pi_x\}_{x\in\mathcal{X}}$ exist, they project onto subspaces of the Hilbert space in which the states $\{\sigma_x\}$ exist, and these projectors and the ensemble satisfy the following conditions:*

$$\operatorname{Tr}\{\sigma_x \Pi\} \geq 1-\epsilon \tag{16.13}$$

$$\operatorname{Tr}\{\sigma_x \Pi_x\} \geq 1-\epsilon \tag{16.14}$$

$$\operatorname{Tr}\{\Pi\} \leq D \tag{16.15}$$

$$\Pi_x \sigma_x \Pi_x \leq \frac{1}{d}\Pi_x \tag{16.16}$$

Suppose that $\mathcal{M}$ is a set of size $|\mathcal{M}|$ with elements m. Let a random covering code $\mathcal{C} \equiv \{C_m\}_{m\in\mathcal{M}}$ consist of random codewords C_m where the codewords C_m are chosen according to the distribution $p_X(x)$ and give rise to a fake ensemble $\left\{\frac{1}{|\mathcal{M}|}, \sigma_{C_m}\right\}_{m\in\mathcal{M}}$. Then there is a high probability that the obfuscation error $o_e(\mathcal{C})$ of the random covering code $\mathcal{C}$ is small:

$$\Pr_{\mathcal{C}}\left\{o_e(\mathcal{C}) \leq \epsilon + 4\sqrt{\epsilon} + 24\sqrt[4]{\epsilon}\right\} \geq 1 - 2D\exp\left(-\frac{\epsilon^3}{4\ln 2}\frac{|\mathcal{M}|d}{D}\right), \tag{16.17}$$

when $|\mathcal{M}| \gg d/D$. Thus it is highly likely that the expected density operator of the fake ensemble $\left\{\frac{1}{|\mathcal{M}|}, \sigma_{C_m}\right\}_{m\in\mathcal{M}}$ is indistinguishable from the expected density operator of the original ensemble $\{p_X(x), \sigma_x\}_{x\in\mathcal{X}}$. It is in this sense that the fake ensemble $\left\{\frac{1}{|\mathcal{M}|}, \sigma_{C_m}\right\}_{m\in\mathcal{M}}$ "covers" the original ensemble $\{p_X(x), \sigma_x\}_{x\in\mathcal{X}}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

16.3 Proof of the Covering Lemma

Before giving the proof of the Covering Lemma, we first state the Operator Chernoff Bound that is useful in proving the Covering Lemma. The Operator Chernoff Bound is a theorem from the theory of large deviations and essentially states that the sample average of a large number of IID random operators is close to the expectation of these random operators (with some constraints on the random operators). The full proof of this lemma appears in Appendix A.

Lemma 16.3.1 (Operator Chernoff Bound)**.** *Let $\xi_1, \ldots, \xi_M$ be M independent and identically distributed random variables with values in the algebra $\mathcal{B}(\mathcal{H})$ of bounded linear operators on some Hilbert space $\mathcal{H}$. Each ξ_m has all of its eigenvalues between the null operator 0 and the identity operator I:*

$$\forall m \in [M] : 0 \leq \xi_m \leq I. \tag{16.18}$$

Let $\overline{\xi}$ denote the sample average of the M random variables:

$$\overline{\xi} = \frac{1}{M} \sum_{m=1}^{M} \xi_m. \tag{16.19}$$

Suppose that the expectation $\mathbb{E}_\xi\{\xi_m\} \equiv \mu$ of each operator ξ_m exceeds the identity operator scaled by a number $a \in (0, 1)$:

$$\mu \geq aI. \tag{16.20}$$

Then for every η where $0 < \eta < 1/2$ and $(1 + \eta)a \leq 1$, we can bound the probability that the sample average $\overline{\xi}$ lies inside the operator interval $[(1 \pm \eta)\mu]$:

$$\Pr_{\xi}\big\{(1-\eta)\mu \leq \overline{\xi} \leq (1+\eta)\mu\big\} \geq 1 - 2\dim\mathcal{H} \exp\left(-\frac{M\eta^2 a}{4\ln 2}\right). \tag{16.21}$$

Thus it is highly likely that the sample average operator $\overline{\xi}$ becomes close to the true expected operator μ as M becomes large.

The first step of the proof of the Covering Lemma is to construct an alternate ensemble that is close to the original ensemble yet satisfies the conditions of the Operator Chernoff Bound (Lemma 16.3.1). We call this alternate ensemble the "Chernoff ensemble." We then generate a random code, a set of M IID random variables, using the Chernoff ensemble. Call this random code the "Chernoff code." We apply the Operator Chernoff Bound to the Chernoff code to obtain a good bound on the obfuscation error of the Chernoff code. We finally show that the bound holds for a covering code generated by the original ensemble because the original ensemble is close to the Chernoff ensemble in trace distance.

16.3.1 Construction of the Chernoff Ensemble

We first establish a few definitions to construct intermediary ensembles. We then use these intermediary ensembles to construct the Chernoff ensemble. We construct the first "primed"

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

ensemble $\{p_X(x), \sigma'_x\}$ by using the projection operators Π_x to slice out some of the support of the states σ_x:

$$\forall x \quad \sigma'_x \equiv \Pi_x \sigma_x \Pi_x. \tag{16.22}$$

The above "slicing" operation cuts outs any elements of the support of σ_x that are not in the support of Π_x. The expected operator σ' for the first primed ensemble is as follows:

$$\sigma' \equiv \sum_{x \in \mathcal{X}} p_X(x) \sigma'_x. \tag{16.23}$$

We then continue slicing with the projector Π and form the second primed ensemble $\{p_X(x), \sigma''_x\}$ as follows:

$$\forall x \quad \sigma''_x \equiv \Pi \sigma'_x \Pi. \tag{16.24}$$

The expected operator for the second primed ensemble is as follows:

$$\sigma'' \equiv \sum_{x \in \mathcal{X}} p_X(x) \sigma''_x. \tag{16.25}$$

Let $\hat{\Pi}$ be the projector onto the subspace spanned by the eigenvectors of σ'' whose corresponding eigenvalues are greater than ϵ/D. We would expect that this extra slicing does not change the state very much when D is large. We construct states ω_x in the Chernoff ensemble by using the projector $\hat{\Pi}$ to slice out some more elements of the support of the original ensemble:

$$\forall x \quad \omega_x \equiv \hat{\Pi} \sigma''_x \hat{\Pi}. \tag{16.26}$$

The expected operator ω for the Chernoff ensemble is then as follows:

$$\omega \equiv \sum_{x \in \mathcal{X}} p_X(x) \omega_x. \tag{16.27}$$

The Chernoff ensemble satisfies the conditions necessary to apply the Chernoff bound. We wait to apply the Chernoff bound and for now show how to construct a random covering code.

16.3.2 Chernoff Code Construction

We present a Shannon-like random coding argument. We construct a covering code $\mathcal{C}$ at random by independently generating $|\mathcal{M}|$ codewords according to the distribution $p_X(x)$. Let $\mathcal{C} = \{c_m\}_{m \in \mathcal{M}}$ be a collection of the realizations c_m of $|\mathcal{M}|$ independent random variables C_m. Each C_m takes a value c_m in $\mathcal{X}$ with probability $p_X(c_m)$ and represents a codeword in the random code $\mathcal{C}$. This process generates the Chernoff code $\mathcal{C}$ consisting of $|\mathcal{M}|$ quantum states $\{\omega_{c_m}\}_{m \in \mathcal{M}}$. The fake expected operator $\overline{\omega}(\mathcal{C})$ of the states in the Chernoff code is as follows:

$$\overline{\omega}(\mathcal{C}) \equiv \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} \omega_{c_m}, \tag{16.28}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

because we assume that Alice randomizes codewords in the Chernoff code according to a uniform distribution (notice that there is a difference in the distribution that we use to choose the code and the distribution that Alice uses to randomize the codewords). The expectation $\mathbb{E}_{\mathcal{C}}\{\omega_{C_m}\}$ of each operator ω_{C_m} is equal to the expected operator ω because of the way that we constructed the covering code. We can also define codes with respect to the primed ensembles as follows: $\{\sigma_{c_m}\}_{m\in\mathcal{M}}$, $\{\sigma'_{c_m}\}_{m\in\mathcal{M}}$, $\{\sigma''_{c_m}\}_{m\in\mathcal{M}}$. These codes respectively have fake expected operators of the following form:

$$\overline{\sigma}(\mathcal{C}) \equiv \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\sigma_{c_m}, \tag{16.29}$$

$$\overline{\sigma}'(\mathcal{C}) \equiv \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\sigma'_{c_m}, \tag{16.30}$$

$$\overline{\sigma}''(\mathcal{C}) \equiv \frac{1}{|\mathcal{M}|}\sum_{m=1}^{|\mathcal{M}|}\sigma''_{c_m}. \tag{16.31}$$

Applying the Chernoff Bound: We make one final modification before applying the Operator Chernoff Bound. The operators ω_{c_m} are in the operator interval between the null operator 0 and $\frac{1}{d}\hat{\Pi}$:

$$\forall m \in \mathcal{M} : 0 \leq \omega_{c_m} \leq \frac{1}{d}\hat{\Pi}. \tag{16.32}$$

The above statement holds because the operators σ'_x satisfy $\sigma'_x = \Pi_x\sigma_x\Pi_x \leq \frac{1}{d}\Pi_x$ (the fourth condition of the Covering Lemma) and this condition implies the following inequalities:

$$\sigma'_x = \Pi_x\sigma_x\Pi_x \leq \frac{1}{d}\Pi_x \tag{16.33}$$

$$\Rightarrow \Pi\sigma'_x\Pi = \sigma''_x \leq \frac{1}{d}\Pi\Pi_x\Pi \leq \frac{1}{d}\Pi \tag{16.34}$$

$$\Rightarrow \omega_x = \hat{\Pi}\sigma''_x\hat{\Pi} \leq \frac{1}{d}\hat{\Pi}\Pi\Pi\hat{\Pi} \leq \frac{1}{d}\hat{\Pi}. \tag{16.35}$$

Therefore, we consider another set of operators (not necessarily density operators) where we scale each ω_{c_m} by d so that

$$\forall m \in \mathcal{M} : 0 \leq d\omega_{c_m} \leq \hat{\Pi}. \tag{16.36}$$

This code satisfies the conditions of the Operator Chernoff Bound with $a = \epsilon d/D$ and with $\hat{\Pi}$ acting as the identity on the subspace onto which it projects. We can now apply the Operator Chernoff Bound to bound the probability that the sample average $\overline{\omega}$ falls in the operator interval $[(1 \pm \epsilon)\omega]$:

$$\Pr\{(1-\epsilon)\omega \leq \overline{\omega} \leq (1+\epsilon)\omega\} = \Pr\{d(1-\epsilon)\omega \leq d\overline{\omega} \leq d(1+\epsilon)\omega\} \tag{16.37}$$

$$\geq 1 - 2\text{Tr}\left\{\hat{\Pi}\right\}\exp\left(-\frac{|\mathcal{M}|\epsilon^2(\epsilon d/D)}{4\ln 2}\right) \tag{16.38}$$

$$\geq 1 - 2D\exp\left(-\frac{\epsilon^3}{4\ln 2}\frac{|\mathcal{M}|d}{D}\right). \tag{16.39}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

16.3.3 Obfuscation error of the covering code

The random covering code is a set of $|\mathcal{M}|$ quantum states $\{\sigma_{C_m}\}_{m\in\mathcal{M}}$ where the quantum states arise from the original ensemble. Recall that our goal is to show that the obfuscation error of the random covering code $\mathcal{C}$,

$$o_e(\mathcal{C}) = \|\overline{\sigma}(\mathcal{C}) - \sigma\|_1, \tag{16.40}$$

has a high probability of being small.

We now show that the obfuscation error of this random covering code is highly likely to be small, by relating it to the Chernoff ensemble. Our method of proof is simply to exploit the triangle inequality, the Gentle Operator Lemma (Lemma 9.4.2), and (9.58) several times. The triangle inequality gives the following bound for the obfuscation error:

$$\begin{aligned}
&o_e(\mathcal{C}) \\
&= \|\overline{\sigma}(\mathcal{C}) - \sigma\|_1 && (16.41)\\
&= \|\overline{\sigma}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C}) - (\overline{\omega}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})) + (\overline{\omega}(\mathcal{C}) - \omega) + (\omega - \sigma'') - (\sigma - \sigma'')\|_1 && (16.42)\\
&\leq \|\overline{\sigma}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})\|_1 + \|\overline{\omega}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})\|_1 \\
&\quad + \|\overline{\omega}(\mathcal{C}) - \omega\|_1 + \|\omega - \sigma''\|_1 + \|\sigma - \sigma''\|_1 && (16.43)
\end{aligned}$$

We show how to obtain a good bound for each of the above five terms.

First consider the rightmost term $\|\sigma - \sigma''\|_1$. Consider that the projected state $\sigma'_x = \Pi_x \sigma_x \Pi_x$ is close to the original state σ_x by applying (16.14) and the Gentle Operator Lemma:

$$\|\sigma_x - \sigma'_x\|_1 \leq 2\sqrt{\epsilon}. \tag{16.44}$$

Consider that

$$\|\sigma'_x - \sigma''_x\|_1 \leq 2\sqrt{\epsilon + 2\sqrt{\epsilon}} \tag{16.45}$$

because $\sigma''_x = \Pi \sigma'_x \Pi$ and from applying the Gentle Operator Lemma to

$$\begin{aligned}
\operatorname{Tr}\{\Pi\sigma'_x\} &\geq \operatorname{Tr}\{\Pi\sigma_x\} - \|\sigma_x - \sigma'_x\|_1 && (16.46)\\
&\geq 1 - \epsilon - 2\sqrt{\epsilon}, && (16.47)
\end{aligned}$$

where the first inequality follows from Exercise 9.1.7 and the second from (16.13) and (16.44). Then the state σ''_x is close to the original state σ_x for all x because

$$\begin{aligned}
\|\sigma_x - \sigma''_x\|_1 &\leq \|\sigma_x - \sigma'_x\|_1 + \|\sigma'_x - \sigma''_x\|_1 && (16.48)\\
&\leq 2\sqrt{\epsilon} + 2\sqrt{\epsilon + 2\sqrt{\epsilon}}, && (16.49)
\end{aligned}$$

where we first applied the triangle inequality and the bounds from (16.44) and (16.45).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Convexity of the trace distance then gives a bound on $\|\sigma - \sigma''\|_1$:

$$\|\sigma - \sigma''\|_1 = \left\| \sum_{x \in \mathcal{X}} p_X(x)\sigma_x - \sum_{x \in \mathcal{X}} p_X(x)\sigma''_x \right\|_1 \tag{16.50}$$

$$= \left\| \sum_{x \in \mathcal{X}} p_X(x)(\sigma_x - \sigma''_x) \right\|_1 \tag{16.51}$$

$$\leq \sum_{x \in \mathcal{X}} p_X(x) \|(\sigma_x - \sigma''_x)\|_1 \tag{16.52}$$

$$\leq \sum_{x \in \mathcal{X}} p_X(x) \left(2\sqrt{\epsilon} + 2\sqrt{\epsilon + 2\sqrt{\epsilon}} \right) \tag{16.53}$$

$$= 2\sqrt{\epsilon} + 2\sqrt{\epsilon + 2\sqrt{\epsilon}}, \tag{16.54}$$

We now consider the second rightmost term $\|\omega - \sigma''\|_1$. The support of σ'' has dimension less than D by (16.15), the third condition in the Covering Lemma. Therefore, eigenvalues smaller than ϵ/D contribute at most ϵ to $\text{Tr}\{\sigma''\}$. We can bound the trace of ω as follows:

$$\text{Tr}\{\omega\} \geq (1 - \epsilon)\text{Tr}\{\sigma''\} \tag{16.55}$$

$$= (1 - \epsilon)\text{Tr}\left\{ \sum_{x \in \mathcal{X}} p_X(x)\sigma''_x \right\} \tag{16.56}$$

$$= \sum_{x \in \mathcal{X}} p_X(x)(1 - \epsilon)\text{Tr}\{\sigma''_x\} \tag{16.57}$$

$$\geq \left(\sum_{x \in \mathcal{X}} p_X(x) \right) (1 - \epsilon)\left(1 - \epsilon - 2\sqrt{\epsilon}\right) \tag{16.58}$$

$$= (1 - \epsilon)\left(1 - \epsilon - 2\sqrt{\epsilon}\right) \tag{16.59}$$

$$\geq 1 - 2\left(\epsilon + \sqrt{\epsilon}\right), \tag{16.60}$$

where the first inequality applies the above "eigenvalue bounding" argument and the second inequality employs the bound in (16.46). This argument shows that average operator of the Chernoff ensemble almost has trace one. We can then apply the Gentle Operator Lemma to $\text{Tr}\{\omega\} \geq 1 - 2(\epsilon + \sqrt{\epsilon})$ to give

$$\|\omega - \sigma''\|_1 \leq 2\sqrt{2\left(\epsilon + \sqrt{\epsilon}\right)}. \tag{16.61}$$

We now consider the middle term $\|\overline{\omega} - \omega\|_1$. The Chernoff bound gives us a probabilistic estimate and not a deterministic estimate like the other two bounds we have shown above. So we suppose for now that the fake operator $\overline{\omega}$ of the Chernoff code is close to the average operator ω of the Chernoff ensemble:

$$\overline{\omega}(\mathcal{C}) \equiv \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \omega_{c_m} \in [(1 \pm \epsilon)\omega]. \tag{16.62}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

With this assumption, it holds that

$$\|\overline{\omega}(\mathcal{C}) - \omega\|_1 \leq \epsilon, \tag{16.63}$$

by employing Lemma A.0.2 from Appendix A and $\mathrm{Tr}\{\omega\} \leq 1$.

We consider the second leftmost term $\|\overline{\omega}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})\|_1$. The following inequality holds

$$\mathrm{Tr}\{\overline{\omega}(\mathcal{C})\} \geq 1 - 3\epsilon - 2\sqrt{\epsilon}. \tag{16.64}$$

because in (16.60) we showed that

$$\mathrm{Tr}\{\omega\} \geq 1 - 2\big(\epsilon + \sqrt{\epsilon}\big), \tag{16.65}$$

and we use the triangle inequality:

$$\begin{aligned}
\mathrm{Tr}\{\overline{\omega}(\mathcal{C})\} &= \|\overline{\omega}(\mathcal{C})\|_1 && (16.66)\\
&= \|\omega - (\omega - \overline{\omega}(\mathcal{C}))\|_1 && (16.67)\\
&\geq \|\omega\|_1 - \|\omega - \overline{\omega}(\mathcal{C})\|_1 && (16.68)\\
&= \mathrm{Tr}\{\omega\} - \|\omega - \overline{\omega}(\mathcal{C})\|_1 && (16.69)\\
&\geq \big(1 - 2\big(\epsilon + \sqrt{\epsilon}\big)\big) - \epsilon && (16.70)\\
&= 1 - 3\epsilon - 2\sqrt{\epsilon}. && (16.71)
\end{aligned}$$

Apply the Gentle Operator Lemma to $\mathrm{Tr}\{\overline{\omega}(\mathcal{C})\} \geq 1 - 3\epsilon - 2\sqrt{\epsilon}$ to give

$$\|\overline{\omega}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})\|_1 \leq 2\sqrt{3\epsilon + 2\sqrt{\epsilon}}. \tag{16.72}$$

We finally bound the leftmost term $\|\overline{\sigma}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})\|_1$. We can use convexity of trace distance and (16.49) to obtain the following bounds:

$$\begin{aligned}
\|\overline{\sigma}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})\|_1 &\leq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \left\|\sigma_{C_m} - \sigma''_{C_m}\right\|_1 && (16.73)\\
&\leq 2\sqrt{\epsilon} + 2\sqrt{\epsilon + 2\sqrt{\epsilon}}. && (16.74)
\end{aligned}$$

We now combine all of the above bounds with the triangle inequality in order to bound

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

the obfuscation error of the covering code $\mathcal{C}$:

$$
\begin{aligned}
& o_e(\mathcal{C}) \\
& = \|\overline{\sigma}(\mathcal{C}) - \sigma\|_1 && (16.75)\\
& = \|\overline{\sigma}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C}) - (\overline{\omega}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})) + (\overline{\omega}(\mathcal{C}) - \omega) + (\omega - \sigma'') - (\sigma - \sigma'')\|_1 && (16.76)\\
& \leq \|\overline{\sigma}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})\|_1 + \|\overline{\omega}(\mathcal{C}) - \overline{\sigma}''(\mathcal{C})\|_1 \\
& \qquad + \|\overline{\omega}(\mathcal{C}) - \omega\|_1 + \|\omega - \sigma''\|_1 + \|\sigma - \sigma''\|_1 && (16.77)\\
& \leq \left(2\sqrt{\epsilon} + 2\sqrt{\epsilon + 2\sqrt{\epsilon}}\right) + \left(2\sqrt{3\epsilon + 2\sqrt{\epsilon}}\right) + \epsilon \\
& \qquad + \left(2\sqrt{2\left(\epsilon + \sqrt{\epsilon}\right)}\right) + \left(2\sqrt{\epsilon} + 2\sqrt{\epsilon + 2\sqrt{\epsilon}}\right) && (16.78)\\
& = \epsilon + 4\sqrt{\epsilon} + 4\sqrt{\epsilon + 2\sqrt{\epsilon}} + 2\sqrt{3\epsilon + 2\sqrt{\epsilon}} + 2\sqrt{2\left(\epsilon + \sqrt{\epsilon}\right)} && (16.79)\\
& \leq \epsilon + 4\sqrt{\epsilon} + 24\sqrt[4]{\epsilon} && (16.80)
\end{aligned}
$$

Observe from the above that the event that the quantity ϵ bounds the obfuscation error $o_e(\mathcal{C})$ of the Chernoff code with states ω_{C_m} implies the event when the quantity $\epsilon + 4\sqrt{\epsilon} + 24\sqrt[4]{\epsilon}$ bounds the obfuscation error $o_e(\mathcal{C})$ of the original code with states σ_{C_m}. Thus, we can bound the probability of obfuscation error of the covering code by applying the Chernoff bound:

$$
\begin{aligned}
\Pr\left\{o_e(\mathcal{C}, \{\sigma_{C_m}\}) \leq \epsilon + 4\sqrt{\epsilon} + 24\sqrt[4]{\epsilon}\right\} &\geq \Pr\{o_e(\mathcal{C}, \{\omega_{C_m}\}) \leq \epsilon\} && (16.81)\\
&\geq 1 - 2D\exp\left(-\frac{\epsilon^3}{4\ln 2}\frac{|\mathcal{M}|d}{D}\right). && (16.82)
\end{aligned}
$$

This argument shows that it is highly likely that a random covering code is good in the sense that it has a low obfuscation error.

Exercise 16.3.1 Prove that the Covering Lemma holds for the same ensemble and a set of projectors for which the following conditions hold:

$$
\begin{aligned}
\sum_{x\in\mathcal{X}} p_X(x)\mathrm{Tr}\{\sigma_x\Pi\} &\geq 1 - \epsilon, && (16.83)\\
\sum_{x\in\mathcal{X}} p_X(x)\mathrm{Tr}\{\sigma_x\Pi_x\} &\geq 1 - \epsilon, && (16.84)\\
\mathrm{Tr}\{\Pi\} &\leq D, && (16.85)\\
\Pi_x\sigma_x\Pi_x &\leq \frac{1}{d}\Pi_x, && (16.86)
\end{aligned}
$$

Exercise 16.3.2 Show that there exists a particular covering code with the property that the obfuscation error is small.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

16.4 History and Further Reading

Ahlswede and Winter introduced the operator Chernoff bound in the context of quantum identification [7]. Winter *et al.* later applied it to quantum measurement compression [259, 260]. Devetak and Winter applied the Covering Lemma to classical compression with quantum side information [74] and to distilling secret key from quantum states [76]. Devetak [68] and Cai *et al.* [51] applied it to private classical communication over a quantum channel, and Groisman *et al.* applied it to study the destruction of correlations in a bipartite state [117].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Part V

Noiseless Quantum Shannon Theory

CHAPTER 17

Schumacher Compression

One of the fundamental tasks in classical information theory is the compression of information. Given access to many uses of a noiseless classical channel, what is the best that a sender and receiver can make of this resource for compressed data transmission? Shannon's compression theorem demonstrates that the Shannon entropy is the fundamental limit for the compression rate in the IID setting (recall the development in Section 13.4). That is, if one compresses at a rate above the Shannon entropy, then it is possible to recover the compressed data perfectly in the asymptotic limit, and otherwise, it is not possible to do so.[1] This theorem establishes the prominent role of the entropy in Shannon's theory of information.

In the quantum world, it very well could be that one day a sender and a receiver would have many uses of a noiseless quantum channel available,[2] and the sender could use this resource to transmit compressed quantum information. But what exactly does this mean in the quantum setting? A simple model of a quantum information source is an ensemble of quantum states $\{p_X(x), |\psi_x\rangle\}$, i.e., the source outputs the state $|\psi_x\rangle$ with probability $p_X(x)$, and the states $\{|\psi_x\rangle\}$ do not necessarily have to form an orthonormal basis. Let us suppose for the moment that the classical data x is available as well, even though this might not necessarily be the case in practice. A naive strategy for compressing this quantum information source would be to ignore the quantum states coming out, handle the classical data instead, and exploit Shannon's compression protocol from Section 13.4. That is, the sender compresses the sequence x^n emitted from the quantum information source at a rate equal to the Shannon entropy $H(X)$, sends the compressed classical bits over the noiseless quantum channels, the receiver reproduces the classical sequence x^n at his end, and finally reconstructs the sequence $|\psi_{x^n}\rangle$ of quantum states corresponding to the classical sequence x^n.

The above strategy will certainly work, but it makes no use of the fact that the noiseless quantum channels are quantum! It is clear that noiseless quantum channels will be expensive

[1] Technically, we did not prove the converse part of Shannon's data compression theorem, but the converse of this chapter suffices for Shannon's classical theorem as well.

[2] How we hope so! If working, coherent fault-tolerant quantum computers come along one day, they stand to benefit from quantum compression protocols.

in practice, and the above strategy is wasteful in this sense because it could have merely exploited classical channels (channels that cannot preserve superpositions) to achieve the same goals. Schumacher compression is a strategy that makes effective use of noiseless quantum channels to compress a quantum information source down to a rate equal to the von Neumann entropy. This has a great benefit from a practical standpoint—recall from Exercise 11.9.2 that the von Neumann entropy of a quantum information source is strictly lower than the source's Shannon entropy if the states in the ensemble are non-orthogonal. In order to execute the protocol, the sender and receiver simply need to know the density operator $\rho \equiv \sum_x p_X(x)|\psi_x\rangle\langle\psi_x|$ of the source. Furthermore, Schumacher compression is provably optimal in the sense that any protocol that compresses a quantum information source of the above form at a rate below the von Neumann entropy cannot have a vanishing error in the asymptotic limit.

Schumacher compression thus gives an operational interpretation of the von Neumann entropy as the fundamental limit on the rate of quantum data compression. Also, it sets the term "qubit" on a firm foundation in an information-theoretic sense as a measure of the amount of quantum information "contained" in a quantum information source.

We begin this chapter by giving the details of the general information processing task corresponding to quantum data compression. We then prove that the von Neumann entropy is an achievable rate of compression and follow by showing that it is optimal (these two respective parts are the direct coding theorem and converse theorem for quantum data compression). We illustrate how much savings one can gain in quantum data compression by detailing a specific example. The final section of the chapter closes with a presentation of more general forms of Schumacher compression.

17.1 The Information Processing Task

We first overview the general task that any quantum compression protocol attempts to accomplish. Three parameters n, R, and ϵ corresponding to the length of the original quantum data sequence, the rate, and the error, respectively, characterize any such protocol. An $(n, R+\delta, \epsilon)$ quantum compression code consists of four steps: state preparation, encoding, transmission, and decoding. Figure 17.1 depicts a general protocol for quantum compression.

State Preparation. The quantum information source outputs a sequence $|\psi_{x^n}\rangle^{A^n}$ of quantum states according to the ensemble $\{p_X(x), |\psi_x\rangle\}$ where

$$|\psi_{x^n}\rangle^{A^n} \equiv |\psi_{x_1}\rangle^{A_1} \otimes \cdots \otimes |\psi_{x_n}\rangle^{A_n}. \tag{17.1}$$

The density operator, from the perspective of someone ignorant of the classical sequence x^n, is equal to the tensor power state $\rho^{\otimes n}$ where

$$\rho \equiv \sum_x p_X(x)|\psi_x\rangle\langle\psi_x|. \tag{17.2}$$

Also, we can think about the purification of the above density operator. That is, an equivalent mathematical picture is to imagine that the quantum information source produces states

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

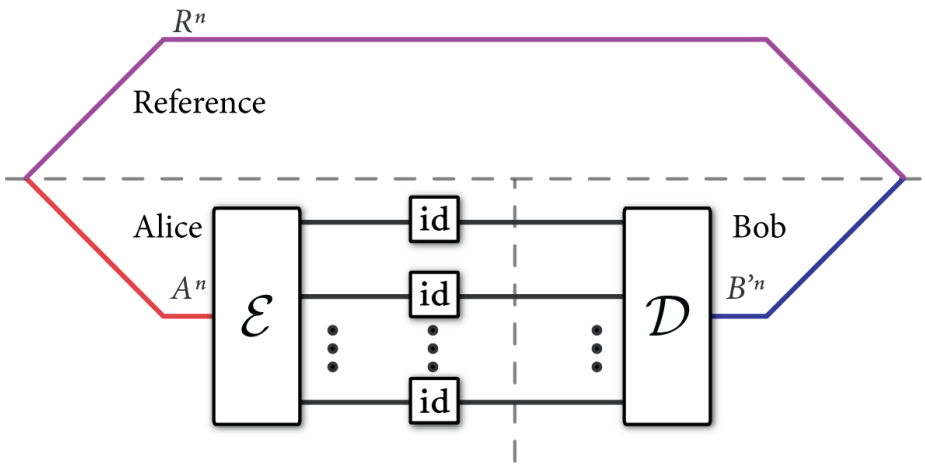

Figure 17.1: The most general protocol for quantum compression. Alice begins with the output of some quantum information source whose density operator is $\rho^{\otimes n}$ on some system A^n. The inaccessible reference system holds the purification of this density operator. She performs some CPTP encoding map $\mathcal{E}$, sends the compressed qubits through 2^{nR} uses of a noiseless quantum channel, and Bob performs some CPTP decoding map $\mathcal{D}$ to decompress the qubits. The scheme is successful if the difference between the initial state and the final state is negligible in the asymptotic limit $n \to \infty$.

of the form

$$|\varphi_\rho\rangle^{RA} \equiv \sum_x \sqrt{p_X(x)}|x\rangle^R|\psi_x\rangle^A, \tag{17.3}$$

where R is the label for an inaccessible reference system (not to be confused with the rate R!). The resulting IID state produced is $(|\varphi_\rho\rangle^{RA})^{\otimes n}$.

Encoding. Alice encodes the systems A^n according to some CPTP compression map $\mathcal{E}^{A^n \to W}$ where W is a quantum system of size 2^{nR}. Recall that R is the rate of compression:

$$R = \frac{1}{n}\log d_W - \delta, \tag{17.4}$$

where d_W is the dimension of system W and δ is an arbitrarily small positive number.

Transmission. Alice transmits the system W to Bob using $n(R+\delta)$ noiseless qubit channels.

Decoding. Bob sends the system W through a decompression map $\mathcal{D}^{W \to \hat{A}^n}$.

The protocol has ϵ error if the compressed and decompressed state is ϵ-close in trace distance to the original state $(|\varphi_\rho\rangle^{RA})^{\otimes n}$:

$$\left\| \left(\varphi_\rho^{RA}\right)^{\otimes n} - (\mathcal{D}^{W \to \hat{A}^n} \circ \mathcal{E}^{A^n \to W})\left(\left(\varphi_\rho^{RA}\right)^{\otimes n}\right) \right\|_1 \leq \epsilon. \tag{17.5}$$

17.2 The Quantum Data Compression Theorem

We say that a quantum compression rate R is *achievable* if there exists an $(n, R+\delta, \epsilon)$ quantum compression code for all $\delta, \epsilon > 0$ and all sufficiently large n. Schumacher's compression

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

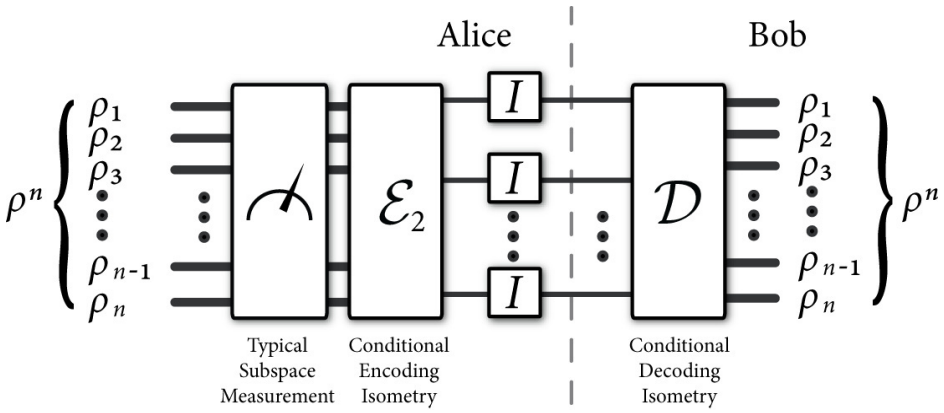

Figure 17.2: Schumacher's compression protocol. Alice begins with many copies of the output of the quantum information source. She performs a measurement onto the typical subspace corresponding to the state ρ and then performs a compression isometry of the typical subspace to a space of size $2^{n[H(\rho)+\delta]}$ qubits. She transmits these compressed qubits over $n[H(\rho)+\delta]$ uses of a noiseless quantum channel. Bob performs the inverse of the isometry to uncompress the qubits. The protocol is successful in the asymptotic limit due to the properties of typical subspaces.

theorem establishes the von Neumann entropy as the fundamental limit on quantum data compression.

Theorem 17.2.1 (Quantum Data Compression)**.** *Suppose that ρ^A is the density operator of the quantum information source. Then the von Neumann entropy $H(A)_\rho$ is the smallest achievable rate R for quantum data compression:*

$$\inf\{R : R \text{ is achievable}\} = H(A)_\rho. \tag{17.6}$$

17.2.1 The Direct Coding Theorem

Schumacher's compression protocol demonstrates that the von Neumann entropy $H(A)_\rho$ is an achievable rate for quantum data compression. It is remarkably similar to Shannon's compression protocol from Section 13.4, but it has some subtle differences that are necessary for the quantum setting. The basic steps of the encoding are to perform a typical subspace measurement and an isometry that compresses the typical subspace. The decoder then performs the inverse of the isometry to decompress the state. The protocol is successful if the typical subspace measurement successfully projects onto the typical subspace, and it fails otherwise. Just like in the classical case, the law of large numbers guarantees that the protocol is successful in the asymptotic limit as $n \to \infty$. Figure 17.2 provides an illustration of the protocol, and we now provide a rigorous argument.

Alice begins with many copies of the state $\left(\varphi_\rho^{RA}\right)^{\otimes n}$. Suppose that the spectral decomposition of ρ is as follows:

$$\rho = \sum_z p_Z(z)|z\rangle\langle z|, \tag{17.7}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $p_Z(z)$ is some probability distribution, and $\{|z\rangle\}$ is some orthonormal basis. Her first step $\mathcal{E}_1^{A^n\to YA^n}$ is to perform a typical subspace measurement of the form in (14.1.4) onto the typical subspace of A^n, where the typical projector is with respect to the density operator ρ. The action of $\mathcal{E}_1^{A^n\to A^nY}$ on a general state σ^{A^n} is

$$\mathcal{E}_1^{A^n\to YA^n}(\sigma^{A^n}) \equiv |0\rangle\langle 0|^Y \otimes \left(I - \Pi_\delta^{A^n}\right)\sigma^{A^n}\left(I - \Pi_\delta^{A^n}\right) + |1\rangle\langle 1|^Y \otimes \Pi_\delta^{A^n}\sigma^{A^n}\Pi_\delta^{A^n}, \quad (17.8)$$

and the classically-correlated flag bit Y indicates whether the typical subspace projection $\Pi_\delta^{A^n}$ is successful or unsuccessful. Recall from the Shannon compression protocol in Section 13.4 that we exploited a function f that mapped from the set of typical sequences to a set of binary sequences $\{0,1\}^{n[H(\rho)+\delta]}$. Now, we can construct an isometry U_f that is a coherent version of this classical function f. It simply maps the orthonormal basis $\{|z^n\rangle^{A^n}\}$ to the basis $\{|f(z^n)\rangle^W\}$:

$$U_f \equiv \sum_{z^n\in T_\delta^{Z^n}} |f(z^n)\rangle^W\langle z^n|^{A^n}, \quad (17.9)$$

where Z is a random variable corresponding to the distribution $p_Z(z)$ so that $T_\delta^{Z^n}$ is its typical set. The above operator is an isometry because the input space is a subspace of size at most $2^{n[H(\rho)+\delta]}$ (recall Property 14.1.2) embedded in a larger space of size 2^n (at least for qubits) and the output space is of size at most $2^{n[H(\rho)+\delta]}$. So her next step $\mathcal{E}_2^{YA^n\to YW}$ is to perform the isometric compression conditional on the flag bit Y being equal to one, and the action of $\mathcal{E}_2^{YA^n\to YW}$ on a general classical-quantum state $\sigma^{YA^n} \equiv |0\rangle\langle 0|^Y\otimes\sigma_0^{A^n} + |1\rangle\langle 1|^Y\otimes\sigma_1^{A^n}$ is as follows:

$$\mathcal{E}_2^{YA^n\to YW}\left(\sigma^{YA^n}\right) \equiv |0\rangle\langle 0|^Y \otimes \mathrm{Tr}\left\{\sigma_0^{A^n}\right\}|e\rangle\langle e|^W + |1\rangle\langle 1|^Y \otimes U_f\sigma_1^{A^n}U_f^\dagger, \quad (17.10)$$

where $|e\rangle^W$ is some error flag orthogonal to all of the states $\{|f(\phi_{x^n})\rangle^W\}_{\phi_{x^n}\in T_\delta^{Z^n}}$. This last step completes the details of her encoder $\mathcal{E}^{A^n\to YW}$, and the action of it on the initial state is

$$\mathcal{E}^{A^n\to YW}(\left(\varphi_\rho^{RA}\right)^{\otimes n}) \equiv (\mathcal{E}_2^{YA^n\to YW}\circ\mathcal{E}_1^{A^n\to YA^n})(\left(\varphi_\rho^{RA}\right)^{\otimes n}). \quad (17.11)$$

Alice then transmits all of the compressed qubits over $n[H(\rho)+\delta]+1$ uses of the noiseless qubit channel.

Bob's decoding $\mathcal{D}^{YW\to A^n}$ performs the inverse of the isometry conditional on the flag bit being equal to one and otherwise maps to some other state $|e\rangle^{A^n}$ outside of the typical subspace. The action of the decoder on some general classical-quantum state $\sigma^{YW} \equiv |0\rangle\langle 0|^Y\otimes\sigma_0^W + |1\rangle\langle 1|^Y\otimes\sigma_1^W$ is

$$\mathcal{D}_1^{YW\to YA^n}\left(\sigma^{YW}\right) \equiv |0\rangle\langle 0|^Y \otimes \mathrm{Tr}\left\{\sigma_0^W\right\}|e\rangle\langle e|^{A^n} + |1\rangle\langle 1|^Y \otimes U_f^\dagger\sigma_1^W U_f. \quad (17.12)$$

The final part of the decoder is to discard the classical flag bit: $\mathcal{D}_2^{YA^n\to A^n} \equiv \mathrm{Tr}_Y\{\cdot\}$. Then $\mathcal{D}^{YW\to A^n} \equiv \mathcal{D}_2^{YA^n\to A^n}\circ\mathcal{D}_1^{YW\to YA^n}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We now can analyze how this protocol performs with respect to our performance criterion in (17.5). Consider the following chain of inequalities:

$$\begin{aligned}
&\left\| \left(\varphi_\rho^{RA}\right)^{\otimes n} - \left(\mathcal{D}^{YW\to A^n} \circ \mathcal{E}^{A^n\to YW}\right)\left(\left(\varphi_\rho^{RA}\right)^{\otimes n}\right) \right\|_1 \\
&= \left\| \text{Tr}_Y\left\{ |1\rangle\langle 1|^Y \otimes \left(\varphi_\rho^{RA}\right)^{\otimes n} \right\} - \left(\mathcal{D}^{YW\to A^n} \circ \mathcal{E}^{A^n\to YW}\right)\left(\left(\varphi_\rho^{RA}\right)^{\otimes n}\right) \right\|_1 && (17.13) \\
&\leq \left\| |1\rangle\langle 1|^Y \otimes \left(\varphi_\rho^{RA}\right)^{\otimes n} - \left(\mathcal{D}_1^{YW\to YA^n} \circ \mathcal{E}^{A^n\to YW}\right)\left(\left(\varphi_\rho^{RA}\right)^{\otimes n}\right) \right\|_1 && (17.14) \\
&= \left\| \begin{array}{c} |1\rangle\langle 1|^Y \otimes \left(\varphi_\rho^{RA}\right)^{\otimes n} - \\ \left(\begin{array}{c} |0\rangle\langle 0|^Y \otimes \text{Tr}\left\{ \left(I - \Pi_\delta^{A^n}\right)\left(\varphi_\rho^{RA}\right)^{\otimes n} \right\} |e\rangle\langle e|^{A^n} \\ + |1\rangle\langle 1|^Y \otimes \Pi_\delta^{A^n}\left(\varphi_\rho^{RA}\right)^{\otimes n}\Pi_\delta^{A^n} \end{array} \right) \end{array} \right\|_1 && (17.15)
\end{aligned}$$

The first equality follows by adding a flag bit $|1\rangle^Y$ to $\left(\varphi_\rho^{RA}\right)^{\otimes n}$ and tracing it out. The first inequality follows from monotonicity of trace distance under the discarding of subsystems (Corollary 9.1.2). The second equality follows by evaluating the map $\mathcal{D}_1^{YW\to A^n} \circ \mathcal{E}^{A^n\to YW}$ on the state $\left(\varphi_\rho^{RA}\right)^{\otimes n}$. Continuing, we have

$$\begin{aligned}
&\leq \left\| |1\rangle\langle 1|^Y \otimes \left(\varphi_\rho^{RA}\right)^{\otimes n} - |1\rangle\langle 1|^Y \otimes \Pi_\delta^{A^n}\left(\varphi_\rho^{RA}\right)^{\otimes n}\Pi_\delta^{A^n} \right\|_1 \\
&\qquad + \left\| |0\rangle\langle 0|^Y \otimes \text{Tr}\left\{ \left(I - \Pi_\delta^{A^n}\right)\left(\varphi_\rho^{RA}\right)^{\otimes n} \right\} |e\rangle\langle e|^{A^n} \right\|_1 && (17.16) \\
&= \left\| \left(\varphi_\rho^{RA}\right)^{\otimes n} - \Pi_\delta^{A^n}\left(\varphi_\rho^{RA}\right)^{\otimes n}\Pi_\delta^{A^n} \right\|_1 + \text{Tr}\left\{ \left(I - \Pi_\delta^{A^n}\right)\left(\varphi_\rho^{RA}\right)^{\otimes n} \right\} && (17.17) \\
&\leq 2\sqrt{\epsilon} + \epsilon && (17.18)
\end{aligned}$$

The first inequality follows from the triangle inequality for trace distance (Lemma 9.1.2). The equality uses the facts $\|\rho \otimes \sigma - \omega \otimes \sigma\|_1 = \|\rho - \omega\|_1 \|\sigma\|_1 = \|\rho - \omega\|_1$ and $\|b\rho\|_1 = |b|\|\rho\|_1$ for some density operators ρ, σ, and ω and a constant b. The final inequality follows from the first property of typical subspaces:

$$\text{Tr}\left\{ \Pi_\delta^{A^n}\left(\varphi_\rho^{RA}\right)^{\otimes n} \right\} = \text{Tr}\left\{ \Pi_\delta^{A^n}\rho^{\otimes n} \right\} \geq 1 - \epsilon, \tag{17.19}$$

and the Gentle Operator Lemma (Lemma 9.4.2).

We remark that it is important for the typical subspace measurement in (17.8) to be implemented as a coherent quantum measurement. That is, the only information that this measurement should learn is whether the state is typical or not. Otherwise, there would be too much disturbance to the quantum information, and the protocol would fail at the desired task of compression. Such precise control on so many qubits is possible in principle, but it is of course rather daunting to implement in practice!

17.2.2 The Converse Theorem

We now prove the converse theorem for quantum data compression by considering the most general compression protocol that meets the success criterion in (17.5) and demonstrating

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

that such an asymptotically error-free protocol should have its rate of compression above the von Neumann entropy of the source. Alice would like to compress a state σ that lives on a Hilbert space A^n. The purification $\phi^{R^n A^n}$ of this state lives on the joint systems A^n and R^n where R^n is the purifying system (again, we should not confuse reference system R^n with rate R). If she can compress any system on A^n and recover it faithfully, then she should be able to do so for the purification of the state. An $(n, R+\delta, \epsilon)$ compression code has the property that it can compress at a rate R with only error ϵ. The quantum data processing is

$$A^n \quad \underrightarrow{\mathcal{E}} \quad W \quad \underrightarrow{\mathcal{D}} \quad \hat{A}^n, \tag{17.20}$$

and the following inequality holds for a successful quantum compression protocol:

$$\left\| \omega^{R^n \hat{A}^n} - \phi^{R^n A^n} \right\|_1 \leq \epsilon, \tag{17.21}$$

where

$$\omega^{R^n \hat{A}^n} \equiv \mathcal{D}(\mathcal{E}(\phi^{R^n A^n})). \tag{17.22}$$

Consider the following chain of inequalities:

$$\begin{aligned} 2nR &= \log_2\left(2^{nR}\right) + \log_2\left(2^{nR}\right) && (17.23)\\ &\geq |H(W)_\omega| + |H(W|R^n)_\omega| && (17.24)\\ &\geq |H(W)_\omega - H(W|R^n)_\omega| && (17.25)\\ &= I(W;R^n)_\omega && (17.26) \end{aligned}$$

The first inequality follows from the fact that both the quantum entropy $H(W)$ and the conditional quantum entropy $H(W|R^n)$ cannot be larger than the logarithm of the dimension of the system W (see Property 11.1.3 and Theorem 11.5.1). The second inequality is the triangle inequality. The second equality is from the definition of quantum mutual information. Continuing, we have

$$\begin{aligned} &\geq I\left(\hat{A}^n; R^n\right)_\omega && (17.27)\\ &\geq I\left(\hat{A}^n; R^n\right)_\phi - n\epsilon' && (17.28)\\ &= I(A^n;R^n)_\phi - n\epsilon' && (17.29)\\ &= H(A^n)_\phi + H(R^n)_\phi - H(A^n R^n)_\phi - n\epsilon' && (17.30)\\ &= 2H(A^n)_\phi - n\epsilon' && (17.31) \end{aligned}$$

The first inequality follows from the quantum data processing inequality (Bob processes W with the decoder to get $\hat{A}^n$). The second inequality follows from applying the Alicki-Fannes' inequality (see Exercise 11.9.7) to the success criterion in (17.21) and setting $\epsilon' \equiv 6\epsilon R + 4H_2(\epsilon)/n$. The first equality follows because the systems $\hat{A}^n$ and A^n are isomorphic (they have the same dimension). The second equality is by the definition of quantum mutual information, and the last equality follows because the entropies of each half of a pure,

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

bipartite state are equal and their joint entropy vanishes. In the case that the state ϕ is an IID state of the form $(|\varphi_\rho\rangle^{RA})^{\otimes n}$ from before, the von Neumann entropy is additive so that

$$H(A^n)_{|\varphi_\rho\rangle^{\otimes n}} = nH(A)_{|\varphi_\rho\rangle}, \tag{17.32}$$

and this shows that the rate R must be greater than the entropy $H(A)$ if the error of the protocol vanishes in the asymptotic limit.

17.3 Quantum Compression Example

We now highlight a particular example where Schumacher compression gives a big savings in compression rates if noiseless quantum channels are available. Suppose that the ensemble is of the following form:

$$\left\{\left(\frac{1}{2}, |0\rangle\right), \left(\frac{1}{2}, |+\rangle\right)\right\}. \tag{17.33}$$

This ensemble is known as the Bennett-92 ensemble because it is useful in Bennett's protocol for quantum key distribution. The naive strategy would be for Alice and Bob to exploit Shannon's compression protocol. That is, Alice would ignore the quantum nature of the states, and supposing that the classical label for them were available, she would encode the classical label. Though, the entropy of the uniform distribution on two states is equal to one bit, and she would have to transmit classical messages at a rate of one bit per channel use.

A far wiser strategy is to employ Schumacher compression. The density operator of the above ensemble is

$$\frac{1}{2}|0\rangle\langle 0| + \frac{1}{2}|+\rangle\langle +|, \tag{17.34}$$

which has the following spectral decomposition:

$$\cos^2(\pi/8)|+'\rangle\langle +'| + \sin^2(\pi/8)|-'\rangle\langle -'|, \tag{17.35}$$

where

$$|+'\rangle \equiv \cos(\pi/8)|0\rangle + \sin(\pi/8)|1\rangle, \tag{17.36}$$

$$|-'\rangle \equiv \sin(\pi/8)|0\rangle - \cos(\pi/8)|1\rangle. \tag{17.37}$$

The binary entropy $H_2(\cos^2(\pi/8))$ of the distribution $\left[\cos^2(\pi/8), \sin^2(\pi/8)\right]$ is approximately equal to

$$0.6009 \text{ qubits}, \tag{17.38}$$

and thus they can save a significant amount in terms of compression rate by employing Schumacher compression. This type of savings will always occur whenever the ensemble includes non-orthogonal quantum states.

Exercise 17.3.1 In the above example, suppose that Alice associates a classical label with the states, so that the ensemble instead is

$$\left\{\left(\frac{1}{2}, |0\rangle\langle 0| \otimes |0\rangle\langle 0|\right), \left(\frac{1}{2}, |1\rangle\langle 1| \otimes |+\rangle\langle +|\right)\right\}. \tag{17.39}$$

Does this help in reducing the amount of qubits she has to transmit to Bob?

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

17.4 Variations on the Schumacher Theme

We can propose several variations on the Schumacher compression theme. For example, suppose that the quantum information source corresponds to the following ensemble instead:

$$\left\{p_X(x), \rho_x^A\right\}, \tag{17.40}$$

where each ρ_x is a mixed state. Then the situation is not as "clear-cut" as in the simpler model for a quantum information source because the techniques exploited in the converse proof do not apply here. Thus, the entropy of the source does not serve as a lower bound on the ultimate compressibility rate.

Let us consider a special example of the above situation. Suppose that the mixed states ρ_x live on orthogonal subspaces, and let $\rho^A = \sum_x p_X(x)\rho_x$ denote the expected density operator of the ensemble. These states are perfectly distinguishable by a measurement whose projectors project onto the different orthogonal subspaces. Alice could then perform this measurement and associate classical labels with each of the states:

$$\rho^{XA} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A. \tag{17.41}$$

Furthermore, she can do this in principle without disturbing the state in any way, and therefore the entropy of the state ρ^{XA} is equivalent to the original entropy of the state ρ^A:

$$H(A)_\rho = H(XA)_\rho. \tag{17.42}$$

Naively applying Schumacher compression to such a source is actually not a great strategy here. The compression rate would be equal to

$$H(XA)_\rho = H(X)_\rho + H(A|X)_\rho, \tag{17.43}$$

and in this case, $H(A|X)_\rho \geq 0$ because the conditioning system is classical. For this case, a much better strategy than Schumacher compression is for Alice to measure the classical variable X, compress it with Shannon compression, and transmit to Bob so that he can reconstruct the quantum states at his end of the channel. The rate of compression here is equal to the Shannon entropy $H(X)$ which is provably lower than $H(XA)_\rho$ for this example.

How should we handle the mixed source case in general? Let's consider the direct coding theorem and the converse theorem. The direct coding theorem for this case is essentially equivalent to Schumacher's protocol for quantum compression—there does not appear to be a better approach in the general case. The density operator of the source is equal to

$$\rho^A = \sum_x p_X(x)\rho_x^A. \tag{17.44}$$

A compression rate $R > H(A)_\rho$ is achievable if we form the typical subspace measurement from the typical subspace projector $\Pi_\delta^{A^n}$ onto the state $\left(\rho^A\right)^{\otimes n}$. Although the direct coding

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

theorem stays the same, the converse theorem changes somewhat. A purification of the above density operator is as follows:

$$|\phi\rangle^{XX'RA} \equiv \sum_x \sqrt{p_X(x)}|x\rangle^X|x\rangle^{X'}|\phi_{\rho_x}\rangle^{RA}, \tag{17.45}$$

where each $|\phi_{\rho_x}\rangle^{RA}$ is a purification of ρ_x^A. So the purifying system is the joint system $XX'R$. Let $\omega^{XX'R\hat{A}}$ be the actual state generated by the protocol:

$$\omega^{XX'R\hat{A}} \equiv \mathcal{D}(\mathcal{E}(\phi^{XX'RA})). \tag{17.46}$$

We can now provide an alternate converse proof:

$$\begin{aligned} nR &= \log_2 d_W && (17.47)\\ &\geq H(W)_\omega && (17.48)\\ &\geq H(W)_\omega - H(W|X)_\omega && (17.49)\\ &= I(W;X)_\omega && (17.50)\\ &\geq I(A;X)_\omega && (17.51)\\ &\geq I(A;X)_\phi - n\epsilon' && (17.52) \end{aligned}$$

The first equality follows from evaluating the logarithm of the dimension of system W. The first inequality follows because the von Neumann entropy is less than the logarithm of the dimension of the system. The second inequality follows because $H(W|X)_\omega \geq 0$—the system X is classical when tracing over X' and R. The second equality follows from the definition of quantum mutual information. The third inequality follows from quantum data processing, and the final follows from applying the Alicki-Fannes' inequality (similar to the way that we did for the converse of the quantum data compression theorem). Tracing over X' and R of $|\phi\rangle^{XX'RA}$ gives the following state

$$\sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A, \tag{17.53}$$

demonstrating that the ultimate lower bound on the compression rate of a mixed source is the Holevo information of the ensemble. The next exercise asks you to verify that ensembles of mixed states on orthogonal subspaces saturate this bound.

Exercise 17.4.1 Show that the Holevo information of an ensemble of mixed states on orthogonal subspaces has its Shannon information equal to its Holevo information. Thus, this is an example of a class of ensembles that meet the above lower bound on compressibility.

17.5 Concluding Remarks

Schumacher compression was the first quantum Shannon-theoretic result discovered and is the simplest one that we encounter in this book. The proof is remarkably similar to the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

proof of Shannon's noiseless coding theorem, with the main difference that we should be more careful in the quantum case not to be learning any more information than necessary when performing measurements. The intuition that we gain for future quantum protocols is that it often suffices to consider only what happens to a high probability subspace rather than the whole space itself if our primary goal is to have a small probability of error in a communication task. In fact, this intuition is the same needed for understanding information processing tasks such as entanglement concentration, classical communication, private classical communication, and quantum communication.

The problem of characterizing the lower and upper bounds for the quantum compression rate of a mixed state quantum information source still remains open, despite considerable efforts in this direction. It is only in special cases, such as the example mentioned in Section 17.4, that we know of a matching lower and upper bound as in Schumacher's original theorem.

17.6 History and Further Reading

Ohya and Petz devised the notion of a typical subspace [200], and Schumacher independently introduced typical subspaces and proved the quantum data compression theorem in Ref. [216]. Jozsa and Schumacher later generalized this proof [168], and Lo further generalized the theorem to mixed state sources [186]. There are other generalizations in Refs. [147, 13]. Several schemes for universal quantum data compression exist [166, 167, 31], in which the sender does not need to have a description of the quantum information source in order to compress its output. There are also practical schemes for quantum data compression discussed in work about quantum Huffman codes [48].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 18

Entanglement Concentration

Entanglement is one of the most useful resources in quantum information processing. If a sender and receiver share noiseless entanglement in the form of maximally entangled states, then Chapter 6 showed how they can teleport quantum bits between each other with the help of classical communication, or they can double the capacity of a noiseless qubit channel for transmitting classical information. We will see further applications in Chapter 20 where they can exploit noiseless entanglement to assist in the transmission of classical or quantum data over a noisy quantum channel.

Given the utility of maximal entanglement, a reasonable question is to ask what a sender and receiver can accomplish if they share pure entangled states that are not maximally entangled. In the quantum Shannon-theoretic setting, we make the further assumption that the sender and receiver can share many copies of these pure entangled states. We find out in this chapter that they can "concentrate" these non-maximally entangled states to maximally entangled ebits, and the optimal rate at which they can do so in the asymptotic limit is equal to the "entropy of entanglement" (the von Neumann entropy of half of one copy of the original state). Entanglement concentration is thus another fundamental task in noiseless quantum Shannon theory, and it gives a different operational interpretation to the von Neumann entropy.

Entanglement concentration is perhaps complementary to Schumacher compression in the sense that it gives a firm quantum information theoretic interpretation of the term "ebit" (just as Schumacher compression did so for the term "qubit"), and it demonstrates how the entropy of entanglement is the unique measure of entanglement for pure bipartite states. Despite the similarity to Schumacher compression in this respect, entanglement concentration is a fundamentally different protocol, and we will see that these two protocols are not interchangeable. That is, exploiting the Schumacher compression protocol for the task of entanglement concentration fails at accomplishing the goal of entanglement concentration, and vice versa.

The technique for proving that the von Neumann entropy is an achievable rate for entanglement concentration exploits the method of types outlined in Sections 13.7 and 14.3 for classical and quantum typicality, respectively (the most important property is Prop-

erty 13.7.5 which states that the exponential of the entropy is a lower bound on the size of a typical type class). In hindsight, it is perhaps surprising that a typical type class is exponentially large in the large n limit (on the same order as the typical set itself), and we soon discover the quantum Shannon-theoretic consequences of this result.

We begin this chapter by discussing a simple example of entanglement concentration for a finite number of copies of a state. Section 18.2 then details the information processing task that entanglement concentration attempts to accomplish, and Section 18.3 proves both the direct coding theorem and the converse theorem for entanglement concentration. We then discuss how common randomness concentration is the closest classical analog of the entanglement concentration protocol. Finally, we discuss the differences between Schumacher compression and entanglement concentration, especially how exploiting one protocol to accomplish the other's information processing task results in a failure of the intended goal.

18.1 An Example of Entanglement Concentration

A simple example illustrates the main idea underlying the concentration of entanglement. Consider the following partially entangled state:

$$|\Phi_\theta\rangle^{AB} \equiv \cos(\theta)|00\rangle^{AB} + \sin(\theta)|11\rangle^{AB}, \tag{18.1}$$

where θ is some parameter such that $0 < \theta < \pi/2$. The Schmidt decomposition (Theorem 3.6.1) guarantees that the above state is the most general form for a pure bipartite entangled state on qubits. Now suppose that Alice and Bob share three copies of the above state. We can rewrite the three copies of the above state with some straightforward algebra:

$$\begin{aligned}
&|\Phi_\theta\rangle^{A_1B_1}|\Phi_\theta\rangle^{A_2B_2}|\Phi_\theta\rangle^{A_3B_3} \\
&= \cos^3(\theta)|000\rangle^A|000\rangle^B + \sin^3(\theta)|111\rangle^A|111\rangle^B \qquad (18.2)\\
&\quad + \cos(\theta)\sin^2(\theta)\Big(|110\rangle^A|110\rangle^B + |101\rangle^A|101\rangle^B + |011\rangle^A|011\rangle^B\Big) \\
&\quad + \cos^2(\theta)\sin(\theta)\Big(|100\rangle^A|100\rangle^B + |010\rangle^A|010\rangle^B + |100\rangle^A|100\rangle^B\Big) \\
&= \cos^3(\theta)|000\rangle^A|000\rangle^B + \sin^3(\theta)|111\rangle^A|111\rangle^B \qquad (18.3)\\
&\quad + \sqrt{3}\cos(\theta)\sin^2(\theta)\frac{1}{\sqrt{3}}\Big(|110\rangle^A|110\rangle^B + |101\rangle^A|101\rangle^B + |011\rangle^A|011\rangle^B\Big) \\
&\quad + \sqrt{3}\cos^2(\theta)\sin(\theta)\frac{1}{\sqrt{3}}\Big(|100\rangle^A|100\rangle^B + |010\rangle^A|010\rangle^B + |100\rangle^A|100\rangle^B\Big),
\end{aligned}$$

where we relabel all of the systems on Alice and Bob's respective sides as $A \equiv A_1A_2A_3$ and $B \equiv B_1B_2B_3$. Observe that the subspace with coefficient $\cos^3(\theta)$ whose states have zero "ones" is one-dimensional. The subspace whose states have three "ones" is also one-dimensional. But the subspace with coefficient $\cos(\theta)\sin^2(\theta)$ whose states have two "ones" is three-dimensional, and the same holds for the subspace whose states each have one "one."

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

A protocol for entanglement concentration in this scenario is then straightforward. Alice performs a projective measurement consisting of the operators Π_0, Π_1, Π_2, Π_3 where

$$\Pi_0 \equiv |000\rangle\langle 000|^A, \tag{18.4}$$

$$\Pi_1 \equiv |001\rangle\langle 001|^A + |010\rangle\langle 010|^A + |100\rangle\langle 100|^A, \tag{18.5}$$

$$\Pi_2 \equiv |110\rangle\langle 110|^A + |101\rangle\langle 101|^A + |011\rangle\langle 011|^A, \tag{18.6}$$

$$\Pi_3 \equiv |111\rangle\langle 111|^A. \tag{18.7}$$

The subscript i of the projection operator Π_i corresponds to the Hamming weight of the basis states in the corresponding subspace. Bob can perform the same "Hamming weight" measurement on his side. With probability $\cos^6(\theta) + \sin^6(\theta)$, the procedure fails because it results in $|000\rangle^A|000\rangle^B$ or $|111\rangle^A|111\rangle^B$ which is not a maximally entangled state. But with probability $3\cos^2(\theta)\sin^4(\theta)$, the state is in the subspace with Hamming weight two, and it has the following form:

$$\frac{1}{\sqrt{3}}\Big(|110\rangle^A|110\rangle^B + |101\rangle^A|101\rangle^B + |011\rangle^A|011\rangle^B\Big), \tag{18.8}$$

and with probability $3\cos^4(\theta)\sin^2(\theta)$, the state is in the subspace with Hamming weight one, and it has the following form:

$$\frac{1}{\sqrt{3}}\Big(|100\rangle^A|100\rangle^B + |010\rangle^A|010\rangle^B + |100\rangle^A|100\rangle^B\Big). \tag{18.9}$$

Alice and Bob can then perform local operations on their respective systems to rotate either of these states to a maximally-entangled state with Schmidt rank three:

$$\frac{1}{\sqrt{3}}\Big(|0\rangle^A|0\rangle^B + |1\rangle^A|1\rangle^B + |2\rangle^A|2\rangle^B\Big). \tag{18.10}$$

The simple protocol outlined above is the basis for the entanglement concentration protocol, but it unfortunately fails with a non-negligible probability in this case. On the other hand, if we allow Alice and Bob to have a potentially infinite number of copies of a pure bipartite entangled state, the probability of failing becomes negligible in the asymptotic limit due to the properties of typicality, and each type class subspace contains an exponentially large maximally entangled state. The proof of the direct coding theorem in Section 18.3.1 makes this intuition precise.

Generalizing the procedure outlined above to an arbitrary number of copies is straightforward. Suppose Alice and Bob share n copies of the partially entangled state $|\Phi_\theta\rangle$. We can then write the state as follows:

$$|\Phi_\theta\rangle^{A^n B^n} = \sum_{k=0}^{n} \sqrt{\binom{n}{k}} \cos^{n-k}(\theta)\sin^k(\theta) \left(\frac{1}{\sqrt{\binom{n}{k}}} \sum_{x\,:\,w(x)=k} |x\rangle^{A^n}|x\rangle^{B^n} \right), \tag{18.11}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $w(x)$ is the Hamming weight of the binary vector x. Alice performs a "Hamming weight" measurement whose projective operators are as follows:

$$\Pi_k = \sum_{x \,:\, w(x)=k} |x\rangle\langle x|^{A^n}, \tag{18.12}$$

and the Schmidt rank of the maximally entangled state that they then share is $\binom{n}{k}$.

We can give a rough analysis of the performance of the above protocol when n becomes large by exploiting Stirling's approximation (we just need a handle on the term $\binom{n}{k}$ for large n). Recall that Stirling's approximation is $n! \approx \sqrt{2\pi n}(n/e)^n$, and this gives

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!n-k!} && (18.13)\\
&\approx \frac{\sqrt{2\pi n}(n/e)^n}{\sqrt{2\pi k}(k/e)^k\sqrt{2\pi(n-k)}((n-k)/e)^{n-k}} && (18.14)\\
&= \sqrt{\frac{n}{2\pi k(n-k)}}\frac{n^n}{(n-k)^{n-k}k^k} && (18.15)\\
&= \text{poly}(n)\left(\frac{n-k}{n}\right)^{-(n-k)}\left(\frac{k}{n}\right)^{-k} && (18.16)\\
&= \text{poly}(n)\; 2^{n[-((n-k)/n)\log((n-k)/n)-(k/n)\log(k/n)]} && (18.17)\\
&= \text{poly}(n)\; 2^{nH_2(k/n)}, && (18.18)
\end{aligned}$$

where H_2 is the binary entropy function in (1.1) and $\text{poly}(n)$ indicates a term at most polynomial in n. When n is large, the exponential term $2^{nH_2(k/n)}$ dominates the polynomial $\sqrt{n/2\pi k(n-k)}$, so that the polynomial term begins to behave merely as a constant. So, the protocol is for Alice to perform a typical subspace measurement with respect to the distribution $\big(\cos^2(\theta), \sin^2(\theta)\big)$, and the state then collapses to the following one with high probability:

$$\frac{1}{\mathcal{N}} \sum_{\substack{k=0\,:\\ \left|k/n-\sin^2(\theta)\right| \,\leq\, \delta,\\ \left|(n-k)/n-\cos^2(\theta)\right| \,\leq\, \delta}}^{n} \sqrt{\binom{n}{k}}\cos^{n-k}(\theta)\sin^k(\theta)\left(\frac{1}{\sqrt{\binom{n}{k}}}\sum_{x \,:\, w(x)=k}|x\rangle^{A^n}|x\rangle^{B^n}\right), \tag{18.19}$$

where $\mathcal{N}$ is an appropriate normalization constant. Alice and Bob then both perform a Hamming weight measurement and the state collapses to a state of the form:

$$\frac{1}{\sqrt{\text{poly}(n)\; 2^{nH_2(k/n)}}}\sum_{x \,:\, w(x)=k}|x\rangle^{A^n}|x\rangle^{B^n}, \tag{18.20}$$

depending on the outcome k of the measurement. The above state is a maximally entangled state with Schmidt rank $\text{poly}(n)\; 2^{nH_2(k/n)}$, and it follows that

$$H_2(k/n) \geq H_2\big(\cos^2(\theta)\big) - \delta, \tag{18.21}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

from the assumption that the state first projects into the typical subspace. Alice and Bob can then perform local operations to rotate this state to approximately $nH_2(\cos^2(\theta))$ ebits. Thus, this procedure concentrates the original non-maximally state to ebits at a rate equal to the entropy of entanglement of the state $|\Phi_\theta\rangle^{AB}$ in (18.1). The above proof is a bit rough, and it applies only to entangled qubit systems in a pure state. The direct coding theorem in Section 18.3.1 generalizes this proof to pure entangled states on d-dimensional systems.

18.2 The Information Processing Task

We first detail the information processing task that entanglement concentration sets out to accomplish. An $(n, E-\delta, \epsilon)$ entanglement concentration protocol consists of just one step of processing. Alice and Bob begin with many copies $(|\varphi\rangle^{AB})^{\otimes n}$ of a pure bipartite, entangled state $|\varphi\rangle^{AB}$. Alice and Bob each then perform local CPTP maps $\mathcal{E}^{A^n\to\hat{A}}$ and $\mathcal{F}^{B^n\to\hat{B}}$ in an attempt to concentrate the original state $(|\varphi\rangle^{AB})^{\otimes n}$ to a maximally entangled state:

$$\omega^{\hat{A}\hat{B}} \equiv \left(\mathcal{E}^{A^n\to\hat{A}} \otimes \mathcal{F}^{B^n\to\hat{B}}\right)\left(\varphi^{A^nB^n}\right). \tag{18.22}$$

The protocol has ϵ error if the final state $\omega^{\hat{A}\hat{B}}$ is ϵ-close to a maximally entangled state $|\Phi\rangle^{\hat{A}\hat{B}}$:

$$\left\|\omega^{\hat{A}\hat{B}} - \Phi^{\hat{A}\hat{B}}\right\|_1 \leq \epsilon, \tag{18.23}$$

where

$$|\Phi\rangle^{\hat{A}\hat{B}} \equiv \frac{1}{\sqrt{D}}\sum_{i=0}^{D-1}|i\rangle^{\hat{A}}|i\rangle^{\hat{B}}, \tag{18.24}$$

and the rate E of ebit extraction is

$$E = \frac{1}{n}\log_2(D) + \delta, \tag{18.25}$$

where δ is some arbitrarily small positive number. We say that a particular rate E of entanglement concentration is *achievable* if there exists an $(n, E-\delta, \epsilon)$ entanglement concentration protocol for all $\epsilon, \delta > 0$ and sufficiently large n. Figure 18.1 displays the operation of a general entanglement concentration protocol.

18.3 The Entanglement Concentration Theorem

We first state the entanglement concentration theorem and then prove it below in two parts (the direct coding theorem and the converse theorem).

Theorem 18.3.1 (Entanglement Concentration)**.** *Suppose that $|\varphi\rangle^{AB}$ is a pure bipartite state that Alice and Bob would like to concentrate. Then the von Neumann entropy $H(A)_\varphi$ is the highest achievable rate E for entanglement concentration:*

$$\sup\{E : E \text{ is achievable}\} = H(A)_\varphi. \tag{18.26}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

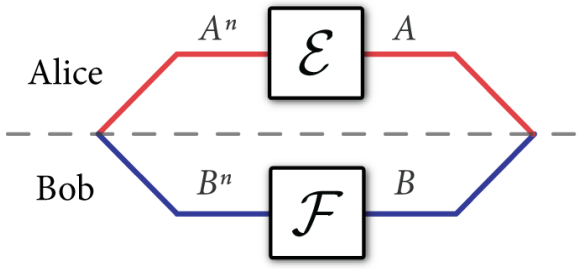

Figure 18.1: The most general protocol for entanglement concentration. Alice and Bob begin with many copies of some pure bipartite state $|\varphi\rangle^{AB}$. They then perform local operations to concentrate this state to a maximally entangled state.

18.3.1 The Direct Coding Theorem

The proof of the direct coding theorem demonstrates the inequality LHS $\geq$ RHS in Theorem 18.3.1. We first outline the technique and then provide a detailed proof. Suppose that Alice and Bob share many copies of a pure, bipartite entangled state $|\varphi\rangle^{AB}$ that has a Schmidt decomposition of the following form:

$$|\varphi\rangle^{AB} = \sum_{x \in \mathcal{X}} \sqrt{p_X(x)} |x\rangle^A |x\rangle^B. \tag{18.27}$$

The fact that the states of Alice and Bob in the above superposition are coordinated via the distribution $p_X(x)$ is what leads to the possibility of performing entanglement concentration (we are assuming that the state above has a Schmidt rank greater than one so that there is some entanglement in it—the protocol given here does not extract any entanglement otherwise). Many copies of the above state admit a type decomposition into different type class subspaces, where each state in a given type class subspace is maximally entangled (just as we observed in the example from Section 18.1). Alice and Bob both first perform a typical subspace measurement onto their n local systems. The fact that their states are coordinated via the distribution $p_X(x)$ implies that they receive the same outcome—the result of the measurement is either typical or atypical for both of them, and the successful typical outcome occurs with high probability in the asymptotic limit. They then perform a type measurement, which in the case of qubits corresponds to a "Hamming weight" measurement. Both of them receive the same outcome for this measurement, and their resulting state is now maximally entangled in the same type class subspace. Furthermore, we know that the size of this type class subspace is larger than $\approx 2^{nH(A)_\varphi}$ in the large n limit because a typical type class has the aforementioned size (Property 13.7.5). They then both perform another projective measurement to bin this subspace into smaller ones in order to guarantee that their maximally entangled state lives in a subspace whose dimension is a power of two (so that they can get ebits) and has the same size for all possible types. Finally, conditional on the type and the bin, they each perform an isometry that rotates their state to $\approx H(A)_\varphi$ ebits. Failure of the above protocol can only occur at two points in this protocol: if the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

outcome of the typical subspace measurement fails or if the second projection onto a slightly smaller subspace fails. Both failures occur with negligible probability in the asymptotic limit.

We now provide a rigorous proof that the above protocol works as claimed. Consider taking many copies of the state $|\varphi\rangle^{AB}$:

$$\left(|\varphi\rangle^{AB}\right)^{\otimes n} = \sum_{x^n \in \mathcal{X}^n} \sqrt{p_{X^n}(x^n)} |x^n\rangle^{A^n} |x^n\rangle^{B^n}. \tag{18.28}$$

We can write the above state in terms of its type decomposition:

$$\begin{aligned} |\varphi\rangle^{A^n B^n} &= \sum_t \sum_{x^n \in T_t} \sqrt{p_{X^n}(x^n)} |x^n\rangle^{A^n} |x^n\rangle^{B^n} && (18.29)\\ &= \sum_t \sqrt{p_{X^n}(x_t^n)} \sum_{x^n \in T_t} |x^n\rangle^{A^n} |x^n\rangle^{B^n} && (18.30)\\ &= \sum_t \sqrt{p_{X^n}(x_t^n) d_t} \frac{1}{\sqrt{d_t}} \sum_{x^n \in T_t} |x^n\rangle^{A^n} |x^n\rangle^{B^n} && (18.31)\\ &= \sum_t \sqrt{p(t)} |\Phi_t\rangle^{A^n B^n}. && (18.32) \end{aligned}$$

The first equality follows by decomposing the state into its different type class subspaces. The next equality follows because $p_{X^n}(x^n)$ is the same for all sequences x^n in the same type class and because the distribution is IID (let x_t^n be some representative sequence of all sequences in the type class T_t). The third equality follows by introducing d_t as the dimension of a type class subspace T_t, and the final equality follows from the definitions

$$\begin{aligned} p(t) &\equiv p_{X^n}(x_t^n) d_t, && (18.33)\\ |\Phi_t\rangle^{A^n B^n} &\equiv \frac{1}{\sqrt{d_t}} \sum_{x^n \in T_t} |x^n\rangle^{A^n} |x^n\rangle^{B^n}. && (18.34) \end{aligned}$$

Observe that the state $|\Phi_t\rangle^{A^n B^n}$ is maximally entangled.

Alice's first action $\mathcal{E}_1^{A^n \to Y A^n}$ is to perform a typical subspace measurement of the form in (14.1.4) onto the typical subspace of A^n, where the typical projector is with respect to the density operator $\rho = \sum_x p_X(x) |x\rangle\langle x|$ (this first step is the same as in Schumacher compression). The action of $\mathcal{E}_1^{A^n \to A^n Y}$ on a general state σ^{A^n} is

$$\begin{aligned} \mathcal{E}_1^{A^n \to Y A^n}(\sigma^{A^n}) \equiv |0\rangle\langle 0|^Y \otimes \left(I - \Pi_\delta^{A^n}\right) \sigma^{A^n} \left(I - \Pi_\delta^{A^n}\right) \\ + |1\rangle\langle 1|^Y \otimes \Pi_\delta^{A^n} \sigma^{A^n} \Pi_\delta^{A^n}, \end{aligned} \tag{18.35}$$

and the classically-correlated flag bit Y indicates whether the typical subspace projection $\Pi_\delta^{A^n}$ is successful or unsuccessful. Conditional on the typical subspace measurement being

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

successful (the flag bit being equal to one), she next performs a type class subspace measurement $\mathcal{E}_2^{YA^n\to YTA^n}$ that places the type in a classical register T. Its action on a general classical-quantum state $\sigma^{YA^n} \equiv |0\rangle\langle 0|^Y \otimes \sigma_0^{A^n} + |1\rangle\langle 1|^Y \otimes \sigma_1^{A^n}$ is as follows:

$$\mathcal{E}_2^{YA^n\to YTA^n}\left(\sigma^{YA^n}\right) = |0\rangle\langle 0|^Y \otimes |e\rangle\langle e|^T \otimes \sigma_0^{A^n} + \sum_t |1\rangle\langle 1|^Y \otimes |t\rangle\langle t|^T \otimes \Pi_t \sigma_1^{A^n} \Pi_t, \quad (18.36)$$

where $\{\Pi_t\}_t$ are the elements of the type class subspace measurement (recall from (14.118) that the typical projector decomposes into a sum of the type class projectors) and $|e\rangle$ is some state that is orthogonal to all of the types $|t\rangle$. Each type class projector Π_t projects onto a subspace of size at least

$$2^{nH(A)_\varphi - \eta(d\delta) - d\log(n+1)}, \quad (18.37)$$

where δ is the typicality parameter, d is the dimension of system A, and $\eta(x)$ is a function such that $\lim_{x\to 0}\eta(x) = 0$. This lower bound follows by exploiting Property 14.3.2. The key observation to make at this point is that Alice and Bob's shared state at this point is a maximally entangled state of the following form:

$$\frac{1}{\sqrt{|T_t|}} \sum_{x^n \in T_t} |x^n\rangle^{A^n} |x^n\rangle^{B^n}. \quad (18.38)$$

This result follows because the distribution $p_{X^n}(x^n)$ becomes uniform when conditioned on a particular type (see the discussion in Section 13.7). Let d_t be the Schmidt rank of the entangled state above.

They now just need to "chop" the above maximally entangled state down to a state of m ebits. Conditional on the flag bit being equal to one and the T register not being equal to $|e\rangle$, Alice and Bob agree beforehand on a partition of their spaces into one of size $(1-\epsilon_1)d_t$ and another of size $\epsilon_1 d_t$ where $\epsilon_1 > 0$. They then further partition the larger space of size $(1-\epsilon_1)d_t$ into Δ bins, each of size 2^m where

$$m \equiv \left\lfloor nH(A)_\varphi - n\eta(d\delta) - d\log(n+1) + \log(1-\epsilon_1) \right\rfloor, \quad (18.39)$$

so that $(1-\epsilon_1)d_t = \Delta\, 2^m$ (they can make the other register of size $\epsilon_1 d_t$ smaller and Δ larger if need be so that Δ is an integer). Conditional on the flag bit being equal to one and the T register not being equal to $|e\rangle$, Alice then performs a projective measurement onto this partitioned type class. The action of this measurement $\mathcal{E}_3^{YTA^n\to YSTA^n}$ on a classical-quantum state of the form

$$\sigma^{YTA^n} \equiv |0\rangle\langle 0|^Y \otimes |e\rangle\langle e|^T \otimes \sigma_0^{A^n} + \sum_t |1\rangle\langle 1|^Y \otimes |t\rangle\langle t|^T \otimes \sigma_t^{A^n} \quad (18.40)$$

is as follows:

$$\mathcal{E}_3^{YTA^n\to YSTA^n}\left(\sigma^{YTA^n}\right) = |0\rangle\langle 0|^Y \otimes |0\rangle\langle 0|^S \otimes |e\rangle\langle e|^T \otimes \sigma_0^{A^n} + \sum_{t,s\,:\,s\neq 0} |1\rangle\langle 1|^Y \otimes |s\rangle\langle s|^S \otimes |t\rangle\langle t|^T \otimes \Pi_{s,t}\sigma_t^{A^n}\Pi_{s,t}, \quad (18.41)$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $\Pi_{0,t}$ is a projector indicating an unsuccessful projection and $\Pi_{s,t}$ is a projector indicating a successful projection onto one of the subspaces of size given in (18.39). Alice's final processing step $\mathcal{E}_4^{YSTA^n \to YST\hat{A}}$ is to perform an isometry $U_t^{A^n \to \hat{A}}$ conditional on the Y register being equal to one, the particular value in the S register, and conditional on the type t, and otherwise trace out A^n and replace it by some state orthogonal to all the binary numbers in the set $\{0,1\}^m$ (indicating failure). The isometry $U_t^{A^n \to \hat{A}}$ is a coherent version of a function g that maps the sequences x^n in the type class T_t' to a binary number in $\{0,1\}^m$:

$$U_t^{A^n \to \hat{A}} \equiv \sum_{x^n \in T_t'} |g(x^n)\rangle\langle x^n|. \tag{18.42}$$

We place a prime on the type class T_t because it is a set slightly smaller than T_t due to the projection in (18.41). Alice then traces out the systems Y, S, and T—let $\mathcal{E}_5^{YST\hat{A} \to \hat{A}}$ denote this action. This last step completes Alice's actions, and let $\mathcal{E}^{A^n \to \hat{A}}$ denote the full action of her "entanglement concentrator":

$$\mathcal{E}^{A^n \to \hat{A}} \equiv \mathcal{E}_5^{YST\hat{A} \to \hat{A}} \circ \mathcal{E}_4^{YSTA^n \to YST\hat{A}} \circ \mathcal{E}_3^{YTA^n \to YSTA^n} \circ \mathcal{E}_2^{YA^n \to YTA^n} \circ \mathcal{E}_1^{A^n \to YA^n}. \tag{18.43}$$

We have outlined all of Alice's steps above, but what should Bob do? It turns out that he should perform *the exact same steps*. Doing the same steps guarantees that he receives the same results at every step because the initial state $|\varphi\rangle^{AB}$ has the special form in (18.27) from the Schmidt decomposition. That is, their classical registers Y, S, and T are perfectly correlated at every step of the concentration protocol. We should also make a point concerning the state after Alice and Bob complete their fourth processing step $\mathcal{E}_4$. Conditional on the Y and S registers being equal to one (which occurs with very high probability in the asymptotic limit), the state on systems $\hat{A}$ and B is equal to a state of m ebits $|\Phi^+\rangle^{\otimes m}$, and so the protocol is successful.

We now perform the error analysis of this protocol. Let $\omega^{\hat{A}\hat{B}}$ denote the state at the end of the protocol:

$$\omega^{\hat{A}\hat{B}} \equiv (\mathcal{E}^{A^n \to \hat{A}} \otimes \mathcal{F}^{B^n \to \hat{B}})(\varphi^{A^n B^n}), \tag{18.44}$$

where $\mathcal{F}$ indicates Bob's steps that are the same as Alice's in (18.43). Consider the following chain of inequalities:

$$\left\| \omega^{\hat{A}\hat{B}} - (\Phi^+)^{\otimes m} \right\|_1 \tag{18.45}$$

$$= \left\| (\mathcal{E}^{A^n \to \hat{A}} \otimes \mathcal{F}^{B^n \to \hat{B}})(\varphi^{A^n B^n}) - (\Phi^+)^{\otimes m} \right\|_1 \tag{18.46}$$

$$\leq \left\| \begin{array}{l} (\mathcal{E}_4 \circ \mathcal{E}_3 \circ \mathcal{E}_2 \circ \mathcal{E}_1) \otimes (\mathcal{F}_4 \circ \mathcal{F}_3 \circ \mathcal{F}_2 \circ \mathcal{F}_1)(\varphi^{A^n B^n}) \\ - \sum_{s \,:\, s \neq 0, t \in \tau_\delta} \frac{1}{\mathcal{N}} \mathrm{Tr}\{\Pi_{s,t} \Pi_t \Pi_\delta^{A^n} \varphi^{A^n B^n}\} |1\rangle\langle 1|^Y \otimes |s,t\rangle\langle s,t|^{ST} \otimes (\Phi^+)^{\otimes m} \end{array} \right\|_1 \tag{18.47}$$

The first equality follows by definition, and the inequality follows from monotonicity of the trace distance under CPTP maps (recall that the last operations $\mathcal{E}_5$ and $\mathcal{F}_5$ are just tracing out). Also, $\mathcal{N}$ is an appropriate normalization constant related to the probability of the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

typical subspace and the probability of projecting correctly onto one of the blocks of size given in (18.39):

$$\mathcal{N} \equiv \sum_{s \,:\, s\neq 0, t\in\tau_\delta} \text{Tr}\{\Pi_{s,t}\Pi_t\Pi_\delta^{A^n}\varphi^{A^nB^n}\} \geq 1-\epsilon-\epsilon_1. \tag{18.48}$$

The first inequality follows from the result of Exercise 4.1.11 (note that Alice and Bob both have access to the registers Y, S, and T because their operations give the same results). Continuing, the last term in (18.47) is bounded as

$$\begin{aligned}
&\leq \left\| \begin{array}{c} \sum_{s \,:\, s\neq 0, t\in\tau_\delta} \text{Tr}\{\Pi_{s,t}\Pi_t\Pi_\delta^{A^n}\varphi^{A^nB^n}\}|1\rangle\langle 1|^Y \otimes |s,t\rangle\langle s,t|^{ST} \otimes (\Phi^+)^{\otimes m} \\ -\sum_{s \,:\, s\neq 0, t\in\tau_\delta} \frac{1}{\mathcal{N}}\text{Tr}\{\Pi_{s,t}\Pi_t\Pi_\delta^{A^n}\varphi^{A^nB^n}\}|1\rangle\langle 1|^Y \otimes |s,t\rangle\langle s,t|^{S} \otimes (\Phi^+)^{\otimes m} \end{array} \right\|_1 \\
&\quad + \left\| \sum_t |1\rangle\langle 1|^Y \otimes |0\rangle\langle 0|^S \otimes |t\rangle\langle t|^T \otimes |ee\rangle\langle ee|^{A^nB^n} \text{Tr}\{\Pi_{0,t}\Pi_t\Pi_\delta^{A^n}(\varphi^{A^nB^n})\} \right\|_1 \\
&\quad + \left\| |0\rangle\langle 0|^Y \otimes \otimes |0\rangle\langle 0|^S \otimes |e\rangle\langle e|^T \otimes |ee\rangle\langle ee|^{A^nB^n} \text{Tr}\{(I-\Pi_\delta^{A^n})\varphi^{A^nB^n}\} \right\|_1 \qquad (18.49) \\
&\leq |\mathcal{N}-1| \left\| \sum_{\substack{s \,:\, s\neq 0, \\ t\in\tau_\delta}} \frac{1}{\mathcal{N}} \text{Tr}\{\Pi_{s,t}\Pi_t\Pi_\delta^{A^n}\varphi^{A^nB^n}\}|1\rangle\langle 1|^Y \otimes |s,t\rangle\langle s,t|^{ST} \otimes \left(\Phi^+\right)^{\otimes m} \right\|_1 \\
&\quad + \epsilon_1 + \epsilon \qquad (18.50) \\
&\leq \epsilon_1 + \epsilon + \epsilon_1 + \epsilon \qquad (18.51) \\
&= 2\epsilon_1 + 2\epsilon. \qquad (18.52)
\end{aligned}$$

The first inequality follows by noting that the successive operations $\mathcal{E}_4 \circ \mathcal{E}_3 \circ \mathcal{E}_2 \circ \mathcal{E}_1$ and those on Bob's side break into three terms: one corresponding to a successful concentration, an error term if the projection onto blocks of size 2^m fails, and another error term if the first typical subspace projection fails. It also follows from an application of the triangle inequality. The second inequality follows by pulling the normalization constant out of the trace distance, by noting that the probability of the projection onto the remainder register is no larger than ϵ_1, and by applying the bound on the typical projection failing. The last few inequalities follow from the bound in (18.48). The rate of the resulting maximally entangled state is then

$$\frac{1}{n}\log(2^m) = \frac{1}{n}\Big\lfloor nH(A)_\varphi - n\eta(d\delta) - d\log(n+1) + \log(1-\epsilon_1)\Big\rfloor. \tag{18.53}$$

This rate becomes asymptotically close to the entropy of entanglement $H(A)_\varphi$ in the limit where $\delta, \epsilon_1 \to 0$ and $n \to \infty$. We have thus proven that LHS $\geq$ RHS in Theorem 18.3.1, and we have shown the following resource inequality:

$$\langle \varphi^{AB} \rangle \geq H(A)_\varphi [qq]. \tag{18.54}$$

That is, beginning with n copies of a pure, bipartite entangled state φ^{AB}, Alice and Bob can extract $nH(A)_\varphi$ ebits from it with negligible error probability in the asymptotic limit.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

18.3.2 The Converse Theorem

We now prove the converse theorem for entanglement concentration, i.e., the inequality LHS $\leq$ RHS in Theorem 18.3.1. Alice and Bob begin with many copies of the pure state $|\varphi\rangle^{AB}$. In the most general protocol given in Figure 18.1, they both perform local CPTP maps $\mathcal{E}^{A^n\to\hat{A}}$ and $\mathcal{F}^{B^n\to\hat{B}}$ to produce the following state:

$$\omega^{\hat{A}\hat{B}} \equiv \left(\mathcal{E}^{A^n\to\hat{A}} \otimes \mathcal{F}^{B^n\to\hat{B}}\right)\left(\varphi^{A^nB^n}\right). \tag{18.55}$$

If the protocol is successful, then the actual state $\omega^{\hat{A}\hat{B}}$ is ϵ-close to the ideal maximally entangled state $\Phi^{\hat{A}\hat{B}}$:

$$\left\|\omega^{\hat{A}\hat{B}} - \Phi^{\hat{A}\hat{B}}\right\|_1 \leq \epsilon. \tag{18.56}$$

Consider the following chain of inequalities:

$$\begin{aligned}
2nE &= 2H(\hat{A})_\Phi && (18.57)\\
&= H(\hat{A})_\Phi + H(\hat{B})_\Phi - H(\hat{A}\hat{B})_\Phi && (18.58)\\
&= I(\hat{A};\hat{B})_\Phi && (18.59)\\
&\leq I(\hat{A};\hat{B})_\omega + n\epsilon' && (18.60)\\
&\leq I(A^n;B^n)_{\varphi^{\otimes n}} + n\epsilon' && (18.61)\\
&= H(A^n)_{\varphi^{\otimes n}} + H(B^n)_{\varphi^{\otimes n}} - H(A^nB^n)_{\varphi^{\otimes n}} + n\epsilon' && (18.62)\\
&= 2nH(A)_\varphi + n\epsilon'. && (18.63)
\end{aligned}$$

The first equality follows because the entropy of entanglement $H(\hat{A})_\Phi$ of a maximally entangled state $\Phi^{\hat{A}\hat{B}}$ is equal to the logarithm of its Schmidt rank. The next equality follows because $H(\hat{B})_\Phi = H(\hat{A})_\Phi$ and $H(\hat{A}\hat{B})_\Phi = 0$ for a pure bipartite entangled state (see Theorem 11.2.1). The third equality follows from the definition of quantum mutual information. The first inequality follows from applying the Alicki-Fannes' inequality for quantum mutual information to (18.56) with $\epsilon' \equiv 6\epsilon\log|A| + 4H_2(\epsilon)/n$ (see Exercise 11.9.7). The second inequality follows from quantum data processing of both A^n and B^n. The final equalities follow from the same arguments as the first two equalities and because the entropy of a tensor product state is additive.

18.4 Common Randomness Concentration

We now briefly discuss common randomness concentration, which is the closest classical analog to entanglement concentration. The goal of this protocol is to extract uniformly distributed bits from non-uniform common randomness. This discussion should help give insight from the classical world into the entanglement concentration protocol. Our discussion merely "hits the high points" of the protocol without being as detailed as in the direct part of the entanglement concentration theorem. Suppose that Alice and Bob begin with a correlated

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

joint probability distribution $p_X(x)\delta(x,y)$ that is not necessarily maximally correlated (if it were maximally correlated, then the distribution $p_X(x)$ would be uniform, and there would be no need for common randomness concentration). Now suppose that they have access to many copies of this distribution:

$$p_{X^n}(x^n)\delta(x^n,y^n). \tag{18.64}$$

The steps in the protocol for common randomness concentration are similar to those in entanglement concentration. The protocol begins with Alice recording whether her sequence x^n is in the typical set $T_\delta^{X^n}$. If the sequence is typical, she keeps it, and the resulting probability distribution is

$$\frac{1}{\Pr\{T_\delta^{X^n}\}}p_{X^n}(x^n)\delta(x^n,y^n), \tag{18.65}$$

where $x^n \in T_\delta^{X^n}$ and $\Pr\{T_\delta^{X^n}\} \geq 1-\epsilon$ follows from the properties of typical sequences. The following non-typical probability distribution occurs with probability ϵ:

$$\frac{p_{X^n}(x^n)\delta(x^n,y^n)}{1-\Pr\{T_\delta^{X^n}\}}, \tag{18.66}$$

and she declares a failure if the sequence is not typical.

Continuing, Alice then determines the type of the probability distribution in (18.65). The key point here (as for entanglement concentration) is that all sequences within the same type class have the same probability of occurring. Thus, conditioned on a particular type, the resulting sequences have a uniform distribution. The distribution resulting from determining the type is

$$\frac{1}{|T_t|}\delta(x^n,y^n), \tag{18.67}$$

where $x^n \in T_t$. The above probability distribution is then a uniform distribution of size at least

$$\frac{1}{(n+1)^{|\mathcal{X}|}}2^{nH(t)}, \tag{18.68}$$

by applying the bound from the proof of Property 13.7.5. This size is slightly larger than the following size

$$\frac{1}{(n+1)^{|\mathcal{X}|}}2^{n(H(X)-\eta(|\mathcal{X}|\delta))}, \tag{18.69}$$

because t is a typical type. Bob performs the exact same steps, and they both then perform local transformations of the data to convert their sequences to uniform random bits. The rate of the resulting uniform bit extraction is then

$$\frac{1}{n}\log\left(\frac{1}{(n+1)^{|\mathcal{X}|}}2^{n(H(X)-\eta(|\mathcal{X}|\delta))}\right) = H(X)-\eta(|\mathcal{X}|\delta)-\frac{1}{n}|\mathcal{X}|\log(n+1). \tag{18.70}$$

This rate becomes asymptotically close to the entropy $H(X)$ in the limit where $\delta \to 0$ and $n \to \infty$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

18.5 Schumacher Compression versus Entanglement Concentration

The tasks of Schumacher compression from the previous chapter and entanglement concentration from the current chapter might seem as if they should be related—Schumacher compression compresses the output of a quantum information source to its fundamental limit of compressibility, and entanglement concentration concentrates entanglement to its fundamental limit. A natural question to ask is whether Alice and Bob can exploit the technique for Schumacher compression to accomplish the information processing task of entanglement concentration, and vice versa, can they exploit entanglement concentration to accomplish Schumacher compression? Interestingly, the answer is "no" to both questions, and in fact, exploiting one protocol to accomplish the other's information processing task performs remarkably poorly. This has to do with the differences between a typical subspace measurement and a type class subspace measurement, which are the building blocks of Schumacher compression and entanglement concentration, respectively.

First, let us consider exploiting the entanglement concentration protocol to accomplish the goal of Schumacher compression. The method of entanglement concentration exploits a type class subspace projection, and it turns out that such a measurement is just too aggressive (disturbs the state too much) to perform well at Schumacher compression. In fact, the protocol for entanglement concentration is so poor at accomplishing the goal of Schumacher compression that the trace distance between the projected state and the original state reaches its maximum value in the asymptotic limit, implying that the two states are perfectly distinguishable! We state this result as the following theorem.

Theorem 18.5.1. *Suppose that the type class subspace measurement projects onto an empirical distribution slightly different from the distribution $p_X(x)$ in the spectral decomposition of a state ρ. Then the trace distance between a concentrated state and the original state approaches its maximum value of two in the asymptotic limit. Furthermore, even if the empirical distribution of the projected type is the same as $p_X(x)$, the projected state never becomes asymptotically close to the tensor power state $\rho^{\otimes n}$.*

Proof. Consider that the first action of entanglement concentration is to perform a typical subspace measurement followed by a type measurement. These actions result in some state of the following form:

$$\frac{1}{\text{Tr}\{\Pi_t \rho^{\otimes n}\}} \Pi_t \rho^{\otimes n} \Pi_t, \tag{18.71}$$

where t is a typical type and the basis for the type class projectors is the eigenbasis of $\rho = \sum_x p_X(x)|x\rangle\langle x|$. Also, consider that we can write the original state $\rho^{\otimes n}$ as follows:

$$\rho^{\otimes n} = \sum_{t'} \Pi_{t'} \rho^{\otimes n} = \sum_{t'} \Pi_{t'} \rho^{\otimes n} \Pi_{t'}. \tag{18.72}$$

The above equality holds because $\sum_{t'} \Pi_{t'} = I$ and because $\rho^{\otimes n}$ commutes with any type projector $\Pi_{t'}$. We would now like to show that the trace distance between the state in

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

(18.71) and $\rho^{\otimes n}$ asymptotically converges to its maximum value of two when $t \neq p_X$, implying that entanglement concentration is a particularly bad method for Schumacher compression. Consider the following chain of inequalities:

$$\begin{aligned}
&\left\| \frac{1}{\text{Tr}\{\Pi_t \rho^{\otimes n}\}} \Pi_t \rho^{\otimes n} \Pi_t - \rho^{\otimes n} \right\|_1 \\
&= \left\| \frac{1}{\text{Tr}\{\Pi_t \rho^{\otimes n}\}} \Pi_t \rho^{\otimes n} \Pi_t - \sum_{t'} \Pi_{t'} \rho^{\otimes n} \Pi_{t'} \right\|_1 && (18.73) \\
&= \left\| \left(\frac{1}{\text{Tr}\{\Pi_t \rho^{\otimes n}\}} - 1 \right) \Pi_t \rho^{\otimes n} \Pi_t - \sum_{t' \neq t} \Pi_{t'} \rho^{\otimes n} \Pi_{t'} \right\|_1 && (18.74) \\
&= \left| \frac{1}{\text{Tr}\{\Pi_t \rho^{\otimes n}\}} - 1 \right| \left\| \Pi_t \rho^{\otimes n} \Pi_t \right\|_1 + \left\| \sum_{t' \neq t} \Pi_{t'} \rho^{\otimes n} \Pi_{t'} \right\|_1 && (18.75) \\
&= \left| \frac{1}{\text{Tr}\{\Pi_t \rho^{\otimes n}\}} - 1 \right| \text{Tr}\{\Pi_t \rho^{\otimes n}\} + \sum_{t' \neq t} \text{Tr}\{\Pi_{t'} \rho^{\otimes n}\} && (18.76) \\
&= \left| 1 - \text{Tr}\{\Pi_t \rho^{\otimes n}\} \right| + \sum_{t'} \text{Tr}\{\Pi_{t'} \rho^{\otimes n}\} - \text{Tr}\{\Pi_t \rho^{\otimes n}\} && (18.77) \\
&= 2\left(1 - \text{Tr}\{\Pi_t \rho^{\otimes n}\}\right) && (18.78) \\
&\geq 2\left(1 - 2^{-nD\left(\frac{t}{n} || p\right)}\right). && (18.79)
\end{aligned}$$

The first equality follows from (18.72). The second equality follows from straightforward algebra. The third equality follows because all of the type class subspaces are orthogonal to each other. The fourth equality follows because the operators $\Pi_{t'} \rho^{\otimes n} \Pi_{t'}$ are positive for all types t'. The last few equalities are straightforward, and the final inequality follows from the bound in Exercise 13.7.1. This shows that the state from the entanglement concentration is a very poor approximation in the asymptotic limit when the type distribution $\frac{t}{n}$ is different from the distribution p from the spectral decomposition of ρ, implying a positive relative entropy $D\left(\frac{t}{n} || p\right)$. On the other hand, suppose that the empirical distribution $\frac{t}{n}$ is the same as the distribution p. Then, we can rewrite $2(1 - \text{Tr}\{\Pi_t \rho^{\otimes n}\})$ as

$$2\left(1 - \text{Tr}\{\Pi_t \rho^{\otimes n}\}\right) = 2 \sum_{t' \neq t} \text{Tr}\{\Pi_t \rho^{\otimes n}\}. \tag{18.80}$$

The resulting expression includes the probability mass from every type class besides t, and the probability mass of the type class t alone can never approach one asymptotically. It is a subspace smaller than the typical subspace because it does not include all of the other typical types, and such a result would thus contradict the optimality of Schumacher compression. Thus, this approximation is also poor in the asymptotic limit. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We also cannot use the technique from Schumacher compression (a typical subspace measurement) to perform entanglement concentration. It seems like it could be possible, given that the eigenvalues of the state resulting a typical subspace measurement are approximately uniform (recall the third property of typical subspaces—Property 14.1.3). That is, suppose that Alice and Bob share many copies of a state $|\phi\rangle^{AB}$ with Schimdt decomposition $|\phi\rangle^{AB} = \sum_x \sqrt{p_X(x)}|x\rangle^A|x\rangle^B$:

$$\left(|\phi\rangle^{AB}\right)^{\otimes n} = \sum_{x^n \in X^n} \sqrt{p_{X^n}(x^n)}|x^n\rangle^{A^n}|x^n\rangle^{B^n}. \tag{18.81}$$

Then a projection onto the typical subspace succeeds with high probability and results in a state of the following form:

$$\frac{1}{\sqrt{\sum_{x^n \in T^{X^n}} p_{X^n}(x^n)}} \sum_{x^n \in T^{X^n}} \sqrt{p_{X^n}(x^n)}|x^n\rangle^{A^n}|x^n\rangle^{B^n}. \tag{18.82}$$

Consider the following maximally entangled state on the typical subspace:

$$\frac{1}{\sqrt{|T^{X^n}|}} \sum_{x^n \in T^{X^n}} |x^n\rangle^{A^n}|x^n\rangle^{B^n}. \tag{18.83}$$

It seems like the state in (18.82) should be approximately close to the maximally entangled state on the typical subspace because the probability amplitudes of the state in (18.82) are all $\approx 2^{-nH(X)}$. But we can provide a simple counterexample to prove that this is not true. Suppose that we have the following two pure, bipartite states:

$$|\phi\rangle^{AB} \equiv \sqrt{p}|00\rangle^{AB} + \sqrt{1-p}|11\rangle^{AB}, \tag{18.84}$$

$$|\psi\rangle^{AB} \equiv \sqrt{1-p}|00\rangle^{AB} + \sqrt{p}|11\rangle^{AB}, \tag{18.85}$$

where p is some real number strictly between zero and one and not equal to $1/2$. Consider that the fidelity between these two states is equal to $2\sqrt{p(1-p)}$, and observe that $2\sqrt{p(1-p)} < 1$ for the values of p given above. Thus, the fidelity between n copies of these states is equal to $(2\sqrt{p(1-p)})^n$ and approaches zero in the asymptotic limit for the values of p given above. Suppose that Alice performs a typical subspace measurement on many copies of the state $|\psi\rangle^{AB}$, and let ψ' denote the resulting state (it is a state of the form in (18.82)). Let Φ denote the maximally entangled state on the typical subspace of $(|\psi\rangle^{AB})^{\otimes n}$. Now suppose for a contradiction that the trace distance between the maximally entangled state Φ and the projected state ψ' becomes small as n becomes large (i.e., suppose that Schumacher compression performs well for the task of entanglement concentration). Also, consider that the typical subspace of $(|\phi\rangle^{AB})^{\otimes n}$ is the same as the typical subspace of $(|\psi\rangle^{AB})^{\otimes n}$ if we employ a typical subspace measurement with respect to entropic typicality (weak typicality). Let ϕ' denote the typically projected state. We can then apply the result of Exercise 9.2.3 twice and the triangle inequality to bound the fidelity between the maximally entangled state Φ and the state $\phi^{\otimes n}$:

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$$F(\phi', \Phi) \leq F\left(\phi', \psi^{\otimes n}\right) + \left\|\Phi - \psi^{\otimes n}\right\|_1 \tag{18.86}$$

$$\leq F\left(\phi^{\otimes n}, \psi^{\otimes n}\right) + \left\|\phi' - \phi^{\otimes n}\right\|_1 + \left\|\psi^{\otimes n} - \psi'\right\|_1 + \left\|\Phi - \psi'\right\|_1 \tag{18.87}$$

$$\leq \left(2\sqrt{p(1-p)}\right)^n + 2\sqrt{\epsilon} + 2\sqrt{\epsilon} + \epsilon. \tag{18.88}$$

The second inequality follows because a typical subspace measurement succeeds with probability $1 - \epsilon$ and the Gentle Measurement Lemma (Lemma 9.4.1) implies that the trace distances $\|\phi' - \phi^{\otimes n}\|_1$ and $\|\psi^{\otimes n} - \psi'\|_1$ are each less than $2\sqrt{\epsilon}$. The bound on $\|\Phi - \psi'\|_1$ follows from the assumption that Schumacher compression is successful at entanglement concentration. Then taking n to be sufficiently large guarantees that the fidelity $F(\phi', \Phi)$ becomes arbitrarily small. But this result contradicts the assumption that Schumacher compression is successful at entanglement concentration because the typical subspace measurement is the same for the state $\phi^{\otimes n}$. That is, we are led to a contradication because $\phi' = \psi'$ when the typical subspace measurement is with respect to entropic typicality. Thus, the trace distance $\|\Phi - \psi'\|_1$ cannot become arbitrarily small for large n, implying that Schumacher compression does not perform well for the task of entanglement concentration.

18.6 Concluding Remarks

Entanglement concentration was one of the earliest discovered protocols in quantum Shannon theory. The protocol exploits one of the fundamental tools of classical information theory (the method of types), but it applies the method in a coherent fashion so that a type class measurement learns only the type and nothing more. The protocol is similar to Schumacher compression in this regard (in that it learns only the necessary information required to execute the protocol and preserves coherent superpositions), and we will continue to see this idea of applying classical techniques in a coherent way in future quantum Shannon-theoretic protocols. For example, the protocol for quantum communication over a quantum channel is a coherent version of a protocol to transmit private classical information over a quantum channel. Despite the similarity of entanglement concentration to Schumacher compression in the aforementioned regard, the protocols are fundamentally different, leading to a failure of the intended information processing task if one protocol is exploited to accomplish the information processing task of the other.

18.7 History and Further Reading

Elias constructed a protocol for randomness concentration in an early paper [90]. Bennett *et al.* offered two different protocols for entanglement concentration (one of which we developed in this chapter) [21]. Nielsen later connected entanglement concentration protocols to the theory of majorization [195]. Lo and Popescu studied entanglement concentration and the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

classical communication cost of the inverse protocol (entanglement dilution) [188, 187]. Hayden and Winter further elaborated on the communication cost of entanglement dilution [135], as did Harrow and Lo [124]. Kaye and Mosca developed practical networks for entanglement concentration [169], and recently, Blume-Kohout *et al.* took this line of research a step further by considering streaming protocols for entanglement concentration [40]. Hayashi and Matsumoto also developed protocols for universal entanglement concentration [129].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Part VI

Noisy Quantum Shannon Theory

Before quantum information theory became an established discipline in its own right, John R. Pierce issued the following quip at the end of his 1973 retrospective article on the history of information theory [204]:

> "I think that I have never met a physicist who understood information theory. I wish that physicists would stop talking about reformulating information theory and would give us a general expression for the capacity of a channel with quantum effects taken into account rather than a number of special cases."

Since the publication of Pierce's article, we have learned much more about quantum mechanics and information theory than he might have imagined at the time, but we have also realized that there is much more to discover. In spite of all that we have learned, we still unfortunately have not been able to address Pierce's concern in the above quote in full generality.

The most basic question that we could ask in quantum Shannon theory (and the one with which Pierce was concerned) is how much classical information can a sender transmit to a receiver by exploiting a quantum channel. We have determined many special cases of quantum channels for which we do know their classical capacities, but we also now know that this most basic question is still wide open in the general case.

What Pierce may not have imagined at the time is that a quantum channel has a much larger variety of capacities than does a classical channel. For example, we might wish to determine the classical capacity of a quantum channel assisted by entanglement shared between the sender and receiver. We have seen that in the simplest of cases, such as the noiseless qubit channel, shared entanglement boosts the classical capacity up to two bits, and we now refer to this phenomenon as the super-dense coding effect (see Chapter 6). Interestingly, the entanglement-assisted capacity of a quantum channel is one of the few scenarios where we can claim to have a complete understanding of the channel's transmission capabilities. From the results regarding the entanglement-assisted capacity, we have learned that shared entanglement is often a "friend" because it tends to simplify results in both quantum Shannon theory and other subfields of quantum information science.

Additionally, we might consider the capacity of a quantum channel for transmitting quantum information. In 1973, it was not even clear what was meant by "quantum information," but we have since been able to formulate what it means, and we have been able to characterize the quantum capacity of a quantum channel. The task of transmitting quantum information over a quantum channel bears some similarities with the task of transmitting private classical information over that channel, where we are concerned with keeping the classical information private from the environment of the channel. This connection has given insight for achieving good rates of quantum communication over a noisy quantum channel, and there is even a certain class of channels for which we already have a good expression for the quantum capacity (the expression being the coherent information of the channel). Though, the problem of determining a good expression for the quantum capacity in the general case is still wide open.

The remaining chapters of the book are an attempt to summarize many items the quantum information community has learned in the past few decades, all of which are an attempt

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

to address Pierce's concern in various ways. The most important open problem in quantum Shannon theory is to find better expressions for these capacities so that we can actually compute them for an arbitrary quantum channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 19

Classical Communication

This chapter begins our exploration of "dynamic" information processing tasks in quantum Shannon theory, where the term "dynamic" indicates that a quantum channel connects a sender to a receiver and their goal is to exploit this resource for communication. We specifically consider the scenario where a sender Alice would like to communicate classical information to a receiver Bob, and the capacity theorem that we prove here is one particular generalization of Shannon's noisy channel coding theorem from classical information theory (overviewed in Section 2.2). In later chapters, we will see other generalizations of Shannon's theorem, depending on what resources are available to assist their communication or depending on whether they are trying to communicate classical or quantum information. For this reason and others, quantum Shannon theory is quite a bit richer than classical information theory.

The naive approach to communicate classical information over a quantum channel is for Alice and Bob simply to mimic the approach used in Shannon's noisy channel coding theorem. That is, they select a random classical code according to some distribution $p_X(x)$, and Bob performs individual measurements of the outputs of a noisy quantum channel according to some POVM. The POVM at the output induces some conditional probability distribution $p_{Y|X}(y|x)$, which we can in turn think of as an induced noisy classical channel. The classical mutual information $I(X;Y)$ of this channel is an achievable rate for communication, and the best strategy for Alice and Bob is to optimize the mutual information over all of Alice's inputs to the channel and over all measurements that Bob could perform at the output. The resulting quantity is equivalent to Bob's optimized accessible information, which we previously discussed in Section 10.8.2.

If the aforementioned coding strategy were optimal, then there would not be anything much interesting to say for the information processing task of classical communication (in fact, there would not be any need for all of the tools we developed in Chapters 14 and 15!). This is perhaps one first clue that the above strategy is not necessarily optimal. Furthermore, we know from Chapter 11 that the Holevo information is an upper bound to the accessible information, and this bound might prompt us to wonder if it is also an achievable rate for classical communication, given that the accessible information is achievable.

The main theorem of this chapter is the classical capacity theorem (also known as the Holevo-Schumacher-Westmoreland theorem), and it states that the Holevo information of a quantum channel is an achievable rate for classical communication. The Holevo information is easier to manipulate mathematically than is the accessible information. The proof of its achievability demonstrates that the aforementioned strategy is not optimal, and the proof also shows how performing collective measurements over all of the channel outputs allows the sender and receiver to achieve the Holevo information as a rate for classical communication. Thus, this strategy fundamentally makes use of quantum-mechanical effects at the decoder and suggests that such an approach is necessary to achieve the Holevo information. Although this strategy exploits collective measurements at the decoder, it does not make use of entangled states at the encoder. That is, the sender could input quantum states that are entangled across all of the channel inputs, and this encoder entanglement might potentially increase classical communication rates.

One major drawback of the classical capacity theorem (also the case for many other results in quantum Shannon theory) is that it only demonstrates that the Holevo information is an achievable rate for classical communication—the converse theorem is a "multi-letter" converse, meaning that it might be necessary in the general case to evaluate the Holevo information over a potentially infinite number of uses of the channel. The multi-letter nature of the capacity theorem implies that the optimization task for general channels is intractable and thus further implies that we know very little about the actual classical capacity of general quantum channels. Now, there are many natural quantum channels such as the depolarizing channel and the dephasing channel for which the classical capacity is known (the Holevo information becomes "single-letter" for these channels), and these results imply that we have a complete understanding of the classical information transmission capabilities of these channels. All of these results have to do with the additivity of the Holevo information of a quantum channel, which we studied previously in Chapter 12.

We mentioned that the Holevo-Schumacher-Westmoreland coding strategy does not make use of entangled inputs at the encoder. But a natural question is to wonder whether entanglement at the encoder could boost classical information transmission rates, given that it is a resource for many quantum protocols. This question was known as the additivity conjecture and went unsolved for many years, but recently Hastings offered a proof that entangled inputs can increase communication rates for certain channels. Thus, for these channels, the single-letter Holevo information is not the proper characterization of classical capacity (though, this is not to say that there could be some alternate characterization of the classical capacity other than the Holevo information which would be single-letter). These recent results demonstrate that we still know little about classical communication in the general case and furthermore that quantum Shannon theory is an active area of research.

We structure this chapter as follows. We first discuss the aforementioned naive strategy in detail, so that we can understand the difference between it and the Holevo-Schumacher-Westmoreland strategy. Section 19.2 describes the steps needed in any protocol for classical communication over a quantum channel. Section 19.3 provides a statement of the classical capacity theorem, and its two subsections prove the corresponding direct coding theorem and

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

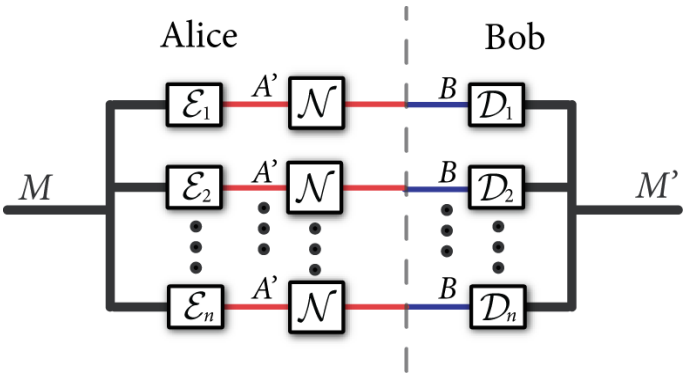

Figure 19.1: The most naive strategy for Alice and Bob to communicate classical information over many independent uses of a quantum channel. Alice wishes to send some message M and selects some tensor product state to input to the channel conditional on the message M. She transmits the codeword over the channel, and Bob then receives a noisy version of it. He performs individual measurements of his quantum systems and produces some estimate M' of the original message M. This scheme is effectively a classical scheme because it makes no use of quantum-mechanical features such as entanglement.

the converse theorem. The direct coding theorem exploits two tools: quantum typicality from Chapter 14 and the packing lemma from Chapter 15. The converse theorem exploits two tools from Chapter 11: continuity of entropies (the Alicki-Fannes' inequality) and the quantum data processing inequality. We then detail how to calculate the classical capacity of several exemplary channels such as entanglement-breaking channels, quantum Hadamard channels, and depolarizing channels—these are channels for which we have a complete understanding of their classical capacity. Finally, we end with a discussion of the recent proof that the Holevo information can be superadditive (that is, entangled inputs at the encoder can enhance classical communication rates for some channels).

19.1 Naive Approach: Product Measurements at the Decoder

We begin by discussing in more detail the most naive strategy that a sender and receiver can exploit for the transmission of classical information over many uses of a quantum channel. Figure 19.1 depicts this naive approach. This first approach mimics certain features of Shannon's classical approach without making any use of quantum-mechanical effects. Alice and Bob agree on a codebook beforehand, where each classical codeword $x^n(m)$ in the codebook corresponds to some message m that Alice wishes to transmit. Alice can exploit some alphabet $\{\rho_x\}$ of density operators to act as input to the quantum channel. That is, the quantum codewords are of the form

$$\rho_{x^n(m)} \equiv \rho_{x_1(m)} \otimes \rho_{x_2(m)} \otimes \cdots \otimes \rho_{x_n(m)}. \tag{19.1}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

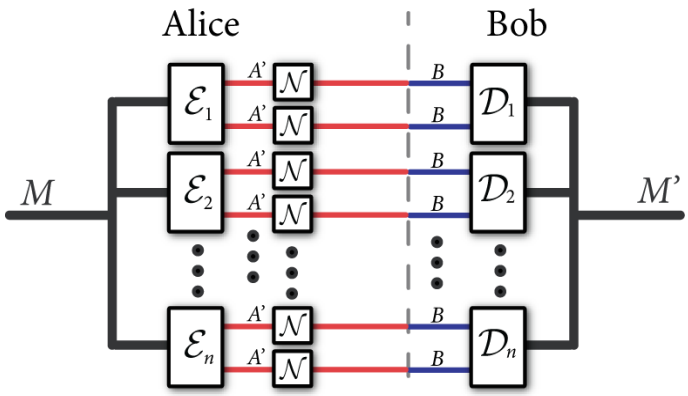

Figure 19.2: A coding strategy that can outperform the previous naive strategy, simply by making use of entanglement at the encoder and decoder.

Bob then performs individual measurements of the outputs of the quantum channel by exploiting some POVM $\{\Lambda_y\}$. This scheme induces the following conditional probability distribution:

$$\begin{aligned}
&p_{Y_1\cdots Y_n|X_1\cdots X_n}(y_1\cdots y_n \mid x_1(m)\cdots x_n(m)) \\
&= \text{Tr}\{\Lambda_{y_1}\otimes\cdots\otimes\Lambda_{y_n}(\mathcal{N}\otimes\cdots\otimes\mathcal{N})(\rho_{x_1(m)}\otimes\cdots\otimes\rho_{x_n(m)})\} &(19.2)\\
&= \text{Tr}\{(\Lambda_{y_1}\otimes\cdots\otimes\Lambda_{y_n})(\mathcal{N}(\rho_{x_1(m)})\otimes\cdots\otimes\mathcal{N}(\rho_{x_n(m)}))\} &(19.3)\\
&= \prod_{i=1}^{n}\text{Tr}\{\Lambda_{y_i}\mathcal{N}(\rho_{x_i(m)})\}, &(19.4)
\end{aligned}$$

which we immediately realize is many independent and identically distributed instances of the following classical channel:

$$p_{Y|X}(y|x) \equiv \text{Tr}\{\mathcal{N}(\rho_x)\Lambda_y\}. \tag{19.5}$$

Thus, if they exploit this scheme, the optimal rate at which they can communicate is equivalent to the following expression:

$$I_{\text{acc}}(\mathcal{N}) \equiv \max_{\{p_X(x),\rho_x,\Lambda\}} I(X;Y), \tag{19.6}$$

where the maximization of the classical mutual information is over all input distributions, all input density operators, and all POVMs that Bob could perform at the output of the channel. This information quantity is known as the accessible information of the channel.

The above strategy is not necessarily an optimal strategy if the channel is truly a quantum channel—it does not make use of any quantum effects such as entanglement. A first simple modification of the protocol to allow for such effects would be to consider coding for the tensor product channel $\mathcal{N}\otimes\mathcal{N}$ rather than the original channel. The input states would be

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

entangled across two channel uses, and the output measurements would be over two channel outputs at a time. In this way, they would be exploiting entangled states at the encoder and collective measurements at the decoder. Figure 19.2 illustrates the modified protocol, and the rate of classical communication that they can achieve with such a strategy is $\frac{1}{2}I_{\text{acc}}(\mathcal{N}\otimes\mathcal{N})$. This quantity is always at least as large as $I_{\text{acc}}(\mathcal{N})$ because a special case of the strategy for the tensor product channel $\mathcal{N}\otimes\mathcal{N}$ is to choose the distribution $p_X(x)$, the states ρ_x, and the POVM Λ to be tensor products of the ones that maximize $I_{\text{acc}}(\mathcal{N})$. We can then extend this construction inductively by forming codes for the tensor product channel $\mathcal{N}^{\otimes k}$ (where k is a positive integer), and this extended strategy achieves the classical communication rate of $\frac{1}{k}I_{\text{acc}}\left(\mathcal{N}^{\otimes k}\right)$ for any finite k. These results then suggest that the ultimate classical capacity of the channel is the regularization of the accessible information of the channel:

$$I_{\text{reg,acc}}(\mathcal{N}) \equiv \lim_{k\to\infty}\frac{1}{k}I_{\text{acc}}\left(\mathcal{N}^{\otimes k}\right). \tag{19.7}$$

The regularization of the accessible information is intractable for general quantum channels, but the optimization task could simplify immensely if the accessible information is additive (additive in the sense of Chapter 12). In this case, the regularized accessible information $I_{\text{reg,acc}}(\mathcal{N})$ would be equivalent to the accessible information $I_{\text{acc}}(\mathcal{N})$. Though, even if the quantity is additive, the optimization could still be difficult to perform in practice. A simple upper bound on the accessible information is the Holevo information $\chi(\mathcal{N})$ of the channel, defined as

$$\chi(\mathcal{N}) \equiv \max_{\rho} I(X;B), \tag{19.8}$$

where the maximization is over classical-quantum states ρ^{XB} of the following form:

$$\rho^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A'\to B}\left(\psi_x^{A'}\right). \tag{19.9}$$

The Holevo information is a more desirable quantity to characterize classical communication over a quantum channel because it is always an upper bound on the accessible information and because Theorem 12.3.2 states that it is sufficient to consider pure states $\psi_x^{A'}$ at the channel input for maximizing the Holevo information.

Thus, a natural question to ask is whether Alice and Bob can achieve the Holevo information rate, and the main theorem of this chapter states that it is possible to do so. The resulting coding scheme bears some similarities with the techniques in Shannon's noisy channel coding theorem, but the main difference is that the decoding POVM is a collective measurement over all of the channel outputs.

19.2 The Information Processing Task

19.2.1 Classical Communication

We now discuss the most general form of the information processing task and give the criterion for a classical communication rate C to be achievable—i.e., we define an $(n, C-\delta, \epsilon)$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

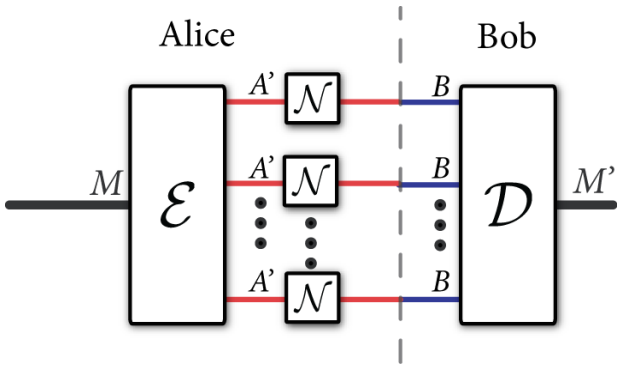

Figure 19.3: The most general protocol for classical communication over a quantum channel. Alice selects some message M and encodes it as a quantum codeword for input to many independent uses of the noisy quantum channel. Bob performs some POVM over all of the channel outputs to determine the message that Alice transmits.

code for classical communication over a quantum channel. Alice begins by selecting some classical message m that she would like to transmit to Bob—she selects from a set of messages $\{1, \ldots, |\mathcal{M}|\}$. Let M denote the random variable corresponding to Alice's choice of message, and let $|\mathcal{M}|$ denote its cardinality. She then prepares some state $\rho_m^{A'^n}$ as input to the many independent uses of the channel—the input systems are n copies of the channel input system A'. She transmits this state over n independent uses of the channel $\mathcal{N}$, and the state at Bob's receiving end is

$$\mathcal{N}^{\otimes n}\left(\rho_m^{A'^n}\right). \tag{19.10}$$

Bob has some decoding POVM $\{\Lambda_m\}$ that he can exploit to determine which message Alice transmits. Figure 19.3 depicts such a general protocol for classical communication over a quantum channel.

Let M' denote the random variable for Bob's estimate of the message. The probability that he determines the correct message m is as follows:

$$\Pr\{M = m \mid M' = m\} = \operatorname{Tr}\left\{\Lambda_m \mathcal{N}^{\otimes n}\left(\rho_m^{A'^n}\right)\right\}, \tag{19.11}$$

and thus the probability of error for a particular message m is

$$p_e(m) \equiv 1 - \Pr\{M = m \mid M' = m\} \tag{19.12}$$

$$= \operatorname{Tr}\left\{(I - \Lambda_m)\mathcal{N}^{\otimes n}\left(\rho_m^{A'^n}\right)\right\}. \tag{19.13}$$

The maximal probability of error for any coding scheme is then

$$p_e^* \equiv \max_{m \in \mathcal{M}} p_e(m). \tag{19.14}$$

The rate C of communication is

$$C \equiv \frac{1}{n}\log_2|\mathcal{M}| + \delta, \tag{19.15}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where δ is some arbitrarily small positive number, and the code has ϵ error if $p_e^* \leq \epsilon$. A rate C of classical communication is *achievable* if there exists an $(n, C - \delta, \epsilon)$ code for all $\delta, \epsilon > 0$ and sufficiently large n.

19.2.2 Common Randomness Generation

A sender and receiver can exploit a quantum channel for the alternate but related task of common randomness generation. Here, they only wish to generate uniform shared randomness of the form:

$$\overline{\Phi}^{MM'} \equiv \frac{1}{|\mathcal{M}|} \sum_{m=1}^{|\mathcal{M}|} |m\rangle\langle m|^M \otimes |m\rangle\langle m|^{M'}. \tag{19.16}$$

Such shared randomness is not particularly useful as a resource, but this viewpoint is helpful for proving the converse theorem of this chapter and later on when we encounter other information processing tasks in quantum Shannon theory. The main point to note is that a noiseless classical bit channel can always generate one bit of noiseless common randomness. Thus, if a quantum channel has a particular capacity for classical communication, it can always achieve the same capacity for common randomness generation. In fact, the capacity for common randomness generation can only be larger than that for classical communication because common randomness is a weaker resource than classical communication. This relationship gives a simple way to bound the capacity for classical communication from above by the capacity for common randomness generation.

The most general protocol for common randomness generation is as follows. Alice begins by locally preparing a state of the form in (19.16). She then performs an encoding map that transforms this state to the following one:

$$\sum_{m=1}^{|\mathcal{M}|} |m\rangle\langle m|^M \otimes \rho_m^{A'^n}, \tag{19.17}$$

and she transmits the A'^n systems over the noisy quantum channel, producing the following state:

$$\sum_{m=1}^{|\mathcal{M}|} |m\rangle\langle m|^M \otimes \mathcal{N}^{\otimes n}\left(\rho_m^{A'^n}\right). \tag{19.18}$$

Bob then performs a quantum instrument on the received systems (exploiting some POVM $\{\Lambda_m\}$), and the resulting state is

$$\omega^{MB^nM'} \equiv \sum_{m=1}^{|\mathcal{M}|} \sum_{m'=1}^{|\mathcal{M}|} |m\rangle\langle m|^M \otimes \sqrt{\Lambda_{m'}}\mathcal{N}^{\otimes n}\left(\rho_m^{A'^n}\right)\sqrt{\Lambda_{m'}} \otimes |m'\rangle\langle m'|^{M'}. \tag{19.19}$$

The state $\omega^{MM'}$ should then be ϵ-close in trace distance to the original state in (19.16) if the protocol is good for common randomness generation:

$$\left\| \overline{\Phi}^{MM'} - \omega^{MM'} \right\|_1 \leq \epsilon. \tag{19.20}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

A rate C for common randomness generation is achievable if there exists an $(n, C - \delta, \epsilon)$ common randomness generation code for all $\delta, \epsilon > 0$ and sufficiently large n.

19.3 The Classical Capacity Theorem

We now state the main theorem of this chapter, the classical capacity theorem.

Theorem 19.3.1 (Holevo-Schumacher-Westmoreland)**.** *The classical capacity of a quantum channel is the supremum over all achievable rates, and one characterization of it is the regularization of the Holevo information of the channel:*

$$\sup\{C \mid C \text{ is achievable}\} = \chi_{reg}(\mathcal{N}), \tag{19.21}$$

where

$$\chi_{reg}(\mathcal{N}) \equiv \lim_{k\to\infty} \frac{1}{k}\chi\left(\mathcal{N}^{\otimes k}\right), \tag{19.22}$$

and the Holevo information $\chi(\mathcal{N})$ of a channel $\mathcal{N}$ is defined in (19.8).

The regularization in the above characterization is a reflection of our ignorance of a better formula for the classical capacity of a quantum channel. The proof of the above theorem in the next two sections demonstrates that the above quantity is indeed equal to the classical capacity, but the regularization implies that the above characterization is intractable for general quantum channels. Though, if the Holevo information of a particular channel is additive (in the sense discussed in Chapter 12), then $\chi_{\text{reg}}(\mathcal{N}) = \chi(\mathcal{N})$, the classical capacity formula simplifies for such a channel, and we can claim to have a complete understanding of the channel's classical transmission capabilities. This "all-or-nothing" situation with capacities is quite common in quantum Shannon theory, and it implies that we still have much remaining to understand about classical information transmission over quantum channels.

The next two sections prove the above capacity theorem in two parts: the direct coding theorem and the converse theorem. The proof of the direct coding theorem demonstrates the inequality LHS $\geq$ RHS in (19.21). That is, it shows that the regularized Holevo information is an achievable rate for classical communication, and it exploits typical and conditionally typical subspaces and the Packing Lemma to do so. The proof of the converse theorem shows the inequality LHS $\leq$ RHS in (19.21). That is, it shows that any protocol with achievable rate C (with vanishing error in the large n limit) should have its rate below the regularized Holevo information. The proof of the converse theorem exploits the aforementioned idea of common randomness generation, continuity of entropy, and the quantum data processing inequality.

19.3.1 The Direct Coding Theorem

We first prove the direct coding theorem. Suppose that a noisy channel $\mathcal{N}$ connects Alice to Bob, and they are allowed access to many independent uses of this quantum channel. Alice

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

can choose some ensemble $\{p_X(x), \rho_x\}$ of states which she can exploit to make a random code for this channel. She selects $|\mathcal{M}|$ codewords $\{x^n(m)\}_{m\in\{1,\ldots,|\mathcal{M}|\}}$ independently according to the following distribution:

$$p'_{X'^n}(x^n) = \begin{cases} \left[\sum_{x^n\in T_\delta^{X^n}} p_{X^n}(x^n)\right]^{-1} p_{X^n}(x^n) & : x^n \in T_\delta^{X^n} \\ 0 & : x^n \notin T_\delta^{X^n} \end{cases}, \tag{19.23}$$

where X'^n is a random variable selected according to the distribution $p'_{X'^n}(x^n)$, $p_{X^n}(x^n) = p_X(x_1)\cdots p_X(x_n)$, and $T_\delta^{X^n}$ denotes the set of strongly typical sequences for the distribution $p_{X^n}(x^n)$ (see Section 13.7). This "pruned" distribution is approximately close to the IID distribution $p_{X^n}(x^n)$ because the probability mass of the typical set is nearly one (the next exercise asks you to make this intuition precise).

Exercise 19.3.1 Prove that the trace distance between the pruned distribution $p'_{X'^n}(x^n)$ and the IID distribution $p_{X^n}(x^n)$ is small for all sufficiently large n:

$$\sum_{x^n\in\mathcal{X}^n} |p'_{X'^n}(x^n) - p_{X^n}(x^n)| \leq 2\epsilon, \tag{19.24}$$

where ϵ is an arbitrarily small positive number such that $\Pr\{X^n \in T_\delta^{X^n}\} \geq 1-\epsilon$.

These classical codewords $\{x^n(m)\}_{m\in\{1,\ldots,|\mathcal{M}|\}}$ lead to quantum codewords of the following form:

$$\rho_{x^n(m)} \equiv \rho_{x_1(m)} \otimes \cdots \otimes \rho_{x_n(m)}, \tag{19.25}$$

by exploiting the quantum states in the ensemble $\{p_X(x), \rho_x\}$. Alice then transmits these codewords through the channel, leading to the following tensor product density operators:

$$\begin{aligned} \sigma_{x^n(m)} &\equiv \sigma_{x_1(m)} \otimes \cdots \otimes \sigma_{x_n(m)} & (19.26)\\ &\equiv \mathcal{N}\big(\rho_{x_1(m)}\big) \otimes \cdots \otimes \mathcal{N}\big(\rho_{x_n(m)}\big). & (19.27) \end{aligned}$$

Bob then detects which codeword Alice transmits by exploiting some detection POVM $\{\Lambda_m\}$ that acts on all of the channel outputs.

At this point, we would like to exploit the Packing Lemma (Lemma 15.3.1 from Chapter 15). Recall that four objects are needed to apply the Packing Lemma, and they should satisfy four inequalities. The first object needed is an ensemble from which we can select a code randomly, and the ensemble in our case is $\{p'_{X'^n}(x^n), \sigma_{x^n}\}$. The next object is the expected density operator of this ensemble:

$$\mathbb{E}_{X'^n}\{\sigma_{X'^n}\} = \sum_{x^n\in\mathcal{X}^n} p'_{X'^n}(x^n)\sigma_{x^n}. \tag{19.28}$$

Finally, we need a message subspace projector and a total subspace projector, and we let these respectively be the conditionally typical projector $\Pi_\delta^{B^n|x^n}$ for the state σ_{x^n} and the typical projector $\Pi_\delta^{B^n}$ for the tensor product state $\sigma^{\otimes n}$ where $\sigma \equiv \sum_x p_X(x)\sigma_x$. Intuitively, the tensor product state $\sigma^{\otimes n}$ should be close to the expected state $\mathbb{E}_{X'^n}\{\sigma_{X'^n}\}$, and the next exercise asks you to verify this statement.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 19.3.2 Prove that the trace distance between the expected state $\mathbb{E}_{X'^n}\{\sigma_{X'^n}\}$ and the tensor product state $\sigma^{\otimes n}$ is small for all sufficiently large n:

$$\left\|\mathbb{E}_{X'^n}\{\sigma_{X'^n}\} - \sigma^{\otimes n}\right\|_1 \leq 2\epsilon, \tag{19.29}$$

where ϵ is an arbitrarily small positive number such that $\Pr\{X^n \in T_\delta^{X^n}\} \geq 1-\epsilon$.

If the four conditions of the Packing Lemma are satisfied (see (15.11-15.14)), then there exists a coding scheme with a detection POVM that has an arbitrarily low maximal probability of error as long as the number of messages in the code is not too high. We now show how to satisfy these four conditions by exploiting the properties of typical and conditionally typical projectors. The following three conditions follow from the properties of typical subspaces:

$$\operatorname{Tr}\{\sigma_{x^n}^{B^n}\Pi_\delta^{B^n}\} \geq 1-\epsilon, \tag{19.30}$$

$$\operatorname{Tr}\left\{\sigma_{x^n}^{B^n}\Pi_\delta^{B^n|x^n}\right\} \geq 1-\epsilon, \tag{19.31}$$

$$\operatorname{Tr}\left\{\Pi_\delta^{B^n|x^n}\right\} \leq 2^{n(H(B|X)+\delta)}. \tag{19.32}$$

The first inequality follows from Property 14.2.7. The second inequality follows from Property 14.2.4, and the third from Property 14.2.5. We leave the proof of the fourth inequality for the Packing Lemma as an exercise.

Exercise 19.3.3 Prove that the following inequality holds

$$\Pi_\delta^{B^n}\mathbb{E}_{X'^n}\{\sigma_{X'^n}\}\Pi_\delta^{B^n} \leq [1-\epsilon]^{-1}2^{-n(H(B)-\delta)}\Pi_\delta^{B^n}. \tag{19.33}$$

(Hint: First show that $\mathbb{E}_{X'^n}\{\sigma_{X'^n}\} \leq [1-\epsilon]^{-1}\sigma^{B^n}$ and then apply the third property of typical subspaces—Property 14.1.3).

With these four conditions holding, it follows from Corollary 15.5.1 (the derandomized version of the Packing Lemma) that there exists a deterministic code and a POVM $\{\Lambda_m\}$ that can detect the transmitted states with arbitrarily low maximal probability of error as long as the size $|\mathcal{M}|$ of the message set is small enough:

$$p_e^* \equiv \max_m \operatorname{Tr}\{(I-\Lambda_m)\mathcal{N}^{\otimes n}(\rho_{x^n(m)})\} \tag{19.34}$$

$$\leq 4\left(\epsilon + 2\sqrt{\epsilon}\right) + 8[1-\epsilon]^{-1}2^{-n(H(B)-H(B|X)-2\delta)}|\mathcal{M}| \tag{19.35}$$

$$= 4\left(\epsilon + 2\sqrt{\epsilon}\right) + 8[1-\epsilon]^{-1}2^{-n(I(X;B)-2\delta)}|\mathcal{M}|. \tag{19.36}$$

So, we can choose the size of the message set to be $|\mathcal{M}| = 2^{n(I(X;B)-3\delta)}$ so that the rate of communication is the Holevo information $I(X;B)$:

$$\frac{1}{n}\log_2|\mathcal{M}| = I(X;B) - 3\delta, \tag{19.37}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and the bound on the maximal probability of error becomes

$$p_e^* \leq 4\left(\epsilon + 2\sqrt{\epsilon}\right) + 8[1-\epsilon]^{-1}2^{-n\delta}. \tag{19.38}$$

Since ϵ is an arbitrary positive number that approaches zero for sufficiently large n and δ is an arbitrarily small positive constant, the maximal probability of error vanishes as n becomes large. Thus, the Holevo information $I(X;B)_\rho$, with respect to the following classical-quantum state

$$\rho^{XB} \equiv \sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}(\rho_x), \tag{19.39}$$

is an achievable rate for the transmission of classical information over $\mathcal{N}$.

Alice and Bob can achieve the Holevo information $\chi(\mathcal{N})$ of the channel $\mathcal{N}$ simply by selecting a random code according to the ensemble $\{p_X(x), \rho_x\}$ that maximizes $I(X;B)_\rho$. Lastly, they can achieve the rate $\frac{1}{k}\chi\left(\mathcal{N}^{\otimes k}\right)$ by coding instead for the tensor product channel $\mathcal{N}^{\otimes k}$, and this last result implies that they can achieve the regularization $\chi_{\text{reg}}(\mathcal{N})$ by making the blocks for which they are coding be arbitrarily large.

We comment more on the role of entanglement at the encoder before moving on to the proof of the converse theorem. First, the above coding scheme for the channel $\mathcal{N}$ does not make use of entangled inputs at the encoder because the codeword states $\rho_{x^n(m)}$ are separable across the channel inputs. It is only when we code for the tensor product channel $\mathcal{N}^{\otimes k}$ that entanglement comes into play. Here, the codeword states are of the form:

$$\rho_{x^n(m)} \equiv \rho_{x_1(m)}^{A'^k} \otimes \cdots \otimes \rho_{x_n(m)}^{A'^k}. \tag{19.40}$$

That is, the states $\rho_{x_i(m)}^{A'^k}$ exist on the Hilbert space of k channel inputs and can be entangled across these k systems. Whether entanglement at the encoder could increase classical communication rates over general quantum channels (whether the regularization in (19.21) is really necessary for the general case) was the subject of much intense work over the past few years, but a recent result has demonstrated the existence of a channel for which exploiting entanglement at the encoder is strictly better than not exploiting entanglement (see Section 19.5).

It is worth re-examining the proof of the Packing Lemma (Lemma 15.3.1) in order to understand better the decoding POVM at the receiving end. The particular decoding POVM elements employed in the Packing Lemma have the following form:

$$\Lambda_m \equiv \left(\sum_{m'=1}^{|\mathcal{M}|} \Gamma_{m'}\right)^{-\frac{1}{2}} \Gamma_m \left(\sum_{m'=1}^{|\mathcal{M}|} \Gamma_{m'}\right)^{-\frac{1}{2}}, \tag{19.41}$$

$$\Gamma_m \equiv \Pi_\delta^{B^n} \Pi_\delta^{B^n|x^n(m)} \Pi_\delta^{B^n}. \tag{19.42}$$

(Simply substitute the conditionally typical projector $\Pi_\delta^{B^n|x^n(m)}$ and the typical projector $\Pi_\delta^{B^n}$ into (15.24).) A POVM with the above elements is known as a "square-root" measurement because of its particular form. We employ such a measurement at the decoder because it

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

has nice analytic properties that allow us to obtain a good bound on the expectation of the average error probability (in particular, we can exploit the operator inequality from Exercise 15.4.1). This measurement is a collective measurement because the conditionally typical projector and the typical projector are both acting on all of the channel outputs, and we construct the square-root measurement from these projectors. Such a decoding POVM is far more exotic than the naive strategy overviewed in Section 19.1 where Bob measures the channel outputs individually—it is for the construction of this decoding POVM and the proof that it is asymptotically good that Holevo, Schumacher, and Westmoreland were given much praise for their work. Though, there is no known way to implement this decoding POVM efficiently, and the original efficiency problems with the decoder in the proof of Shannon's noisy classical channel coding theorem plague the decoders in the quantum world as well.

Exercise 19.3.4 Prove the direct coding theorem of HSW without applying the Packing Lemma (but you can use similar steps as in the Packing Lemma).

Exercise 19.3.5 Show that a measurement with POVM elements of the following form is sufficient to achieve the Holevo information of a quantum channel:

$$\Lambda_m \equiv \left(\sum_{m'=1}^{|\mathcal{M}|} \Pi_\delta^{B^n|x^n(m')} \right)^{-\frac{1}{2}} \Pi_\delta^{B^n|x^n(m)} \left(\sum_{m'=1}^{|\mathcal{M}|} \Pi_\delta^{B^n|x^n(m')} \right)^{-\frac{1}{2}}. \tag{19.43}$$

19.3.2 The Converse Theorem

The second part of the classical capacity theorem is the converse theorem, and we provide a simple proof of it in this section. Suppose that Alice and Bob are trying to accomplish common randomness generation rather than classical communication—the capacity for such a task can only be larger than that for classical communication as we argued before in Section 19.2. Recall that in such a task, Alice first prepares a maximally correlated state $\overline{\Phi}^{MM'}$ so that the rate $C - \delta$ of common randomness generation is equal to $\frac{1}{n}\log_2|M|$. Alice and Bob share a state of the form in (19.19) after encoding, channel transmission, and decoding. We now show that the regularized Holevo information in (19.22) bounds the rate of common randomness generation for any protocol that has vanishing error in the asymptotic limit (the error criterion is in (19.20)). As a result, the regularized Holevo information also upper bounds the capacity for classical communication. Consider the following chain of inequalities:

$$\begin{aligned} n(C-\delta) &= I(M;M')_{\overline{\Phi}} && (19.44)\\ &\leq I(M;M')_{\omega} + n\epsilon' && (19.45)\\ &\leq I(M;B^n)_{\omega} + n\epsilon' && (19.46)\\ &\leq \chi\left(\mathcal{N}^{\otimes n}\right) + n\epsilon'. && (19.47) \end{aligned}$$

The first equality follows because the mutual information of the common randomness state $\overline{\Phi}^{MM'}$ is equal to $n(C-\delta)$ bits. The first inequality follows from the error criterion in

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

(19.20) and by applying the Alicki-Fannes' inequality for quantum mutual information (Exercise 11.9.7) with $\epsilon' \equiv 6\epsilon C + 4H_2(\epsilon)/n$. The second inequality results from the quantum data processing inequality for quantum mutual information (Corollary 11.9.4)—recall that Bob processes the B^n system with a quantum instrument to get the classical system M'. Also, the quantum mutual information is evaluated on a classical-quantum state of the form in (19.18). The final inequality follows because the classical-quantum state in (19.18) has a particular distribution and choice of states, and this choice always leads to a value of the quantum mutual information that cannot be greater than the Holevo information of the tensor product channel $\mathcal{N}^{\otimes n}$.

19.4 Examples of Channels

Observe that the final upper bound in (19.47) on the rate C is the multi-letter Holevo information of the channel. It would be more desirable to have $\chi(\mathcal{N})$ as the upper bound on C rather than $\frac{1}{n}\chi(\mathcal{N}^{\otimes n})$ because the former is simpler, but the optimization problem set out in the latter quantity is simply impossible to compute with finite computational resources. Though, the upper bound in (19.47) is the best known upper bound if we do not know anything else about the structure of the channel, and for this reason, the best known characterization of the classical capacity is the one given in (19.21).

If we know that the Holevo information of the tensor product of a certain channel with itself is additive, then there is no need for the regularization $\chi_{\text{reg}}(\mathcal{N})$, and the characterization in Theorem 19.3.1 reduces to a very good one: the Holevo information $\chi(\mathcal{N})$. There are many examples of channels for which the classical capacity reduces to the Holevo information of the channel, and we detail three such classes of examples in this section: the cq channels, the quantum Hadamard channels, and the quantum depolarizing channels. The proof that demonstrates additivity of the Holevo information for each of these channels depends explicitly on structural properties of each one, and there is unfortunately not much to learn from these proofs in order to say anything about additivity of the Holevo information of general quantum channels. Nevertheless, it is good to have some natural channels for which we can compute their classical capacity, and it is instructive to examine these proofs in detail to understand what it is about each channel that makes their Holevo information additive.

19.4.1 Classical Capacity of Classical-Quantum Channels

Recall from Section 4.4.6 that a quantum channel is a particular kind of entanglement-breaking channel (cq channel) if the action of the channel is equivalent to performing first a complete von Neumann measurement of the input and then preparing a quantum state conditional on the value of the classical variable resulting from the measurement. We have already seen in Section 12.3.1 that the Holevo information of these channels is additive. Additionally, Theorem 12.3.3 states that the Holevo information is a concave function of the input distribution over which we are optimizing for such channels. Thus, computing

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

the classical capacity of cq channels channel can be performed by optimization techniques because its Holevo information is additive.

The Relation to General Channels

We can always exploit the above result regarding cq entanglement-breaking channels to get a reasonable lower bound on the classical capacity of any quantum channel $\mathcal{N}$. The sender Alice can simulate an entanglement-breaking channel by modifying the processing at the input of an arbitrary quantum channel. She can first measure the input to her simulated channel in the basis $\{|x\rangle\langle x|\}$, prepare a state ρ_x conditional on the outcome of the measurement, and subsequently feed this state into the channel $\mathcal{N}$. These actions are equivalent to the following map:

$$\sigma \rightarrow \sum_x \langle x|\sigma|x\rangle \mathcal{N}(\rho_x), \tag{19.48}$$

and the capacity of this simulated channel is equal to

$$I(X;B)_\rho, \tag{19.49}$$

where

$$\rho^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}(\rho_x), \tag{19.50}$$

$$p_X(x) \equiv \langle x|\sigma|x\rangle. \tag{19.51}$$

Of course, Alice has the freedom to prepare whichever state σ she would like to be input to the simulated channel, and she also has the ability to prepare whichever states ρ_x she would like to be conditional on the outcomes of the first measurement, so we should let her maximize the Holevo information over all these inputs. Thus, the capacity of the entanglement-breaking channel composed with the actual channel is equivalent to the Holevo information of the original channel:

$$\max_{p_X(x),\rho_x} I(X;B)_\rho. \tag{19.52}$$

This capacity is also known as the product-state capacity of the channel because it is the capacity achieved by inputting unentangled, separable states at the encoder (Alice can in fact just input product states), and it can be a good lower bound on the true classical capacity of a quantum channel, even if it does not allow for entanglement at the encoder.

19.4.2 Classical Capacity of Quantum Hadamard Channels

Recall from Section 5.2.4 that quantum Hadamard channels are those with a complementary channel that is entanglement-breaking, and this property allows us to prove that the Holevo information of the original channel is additive. Several important natural channels are quantum Hadamard channels. A trivial example is the noiseless qubit channel because Bob could perform a von Neumann measurement of his system and send a constant state

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

to Eve. A less trivial example of a quantum Hadamard channel is a generalized dephasing channel (see Section 5.2.3), though this channel trivially has a maximal classical capacity of $\log_2 d$ bits per channel use because this channel transmits a preferred orthonormal basis without error. A quantum Hadamard channel with a more interesting classical capacity is known as a cloning channel, the channel induced by a universal cloning machine (though we will not discuss this channel in any detail).

Theorem 19.4.1. *The Holevo information of a quantum Hadamard channel $\mathcal{N}_H$ and any other channel $\mathcal{N}$ is additive:*

$$\chi(\mathcal{N}_H \otimes \mathcal{N}) = \chi(\mathcal{N}_H) + \chi(\mathcal{N}). \tag{19.53}$$

Proof. First, recall from Theorem 12.3.2 that it is sufficient to consider ensembles of pure states at the input of the channel when maximizing its Holevo information. That is, we only need to consider classical-quantum states of the following form:

$$\sigma^{XA'} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes |\phi_x\rangle\langle\phi_x|^{A'}, \tag{19.54}$$

where A' is the input to some channel $\mathcal{N}^{A'\to B}$. Let $\omega^{XBE} \equiv U_{\mathcal{N}}^{A'\to BE}(\sigma^{XA'})$ where $U_{\mathcal{N}}^{A'\to BE}$ is an isometric extension of the channel. Thus, the Holevo information of $\mathcal{N}^{A'\to B}$ is equivalent to a different expression:

$$\begin{aligned}
\chi(\mathcal{N}) &\equiv \max_\sigma I(X;B)_\omega && (19.55)\\
&= \max_\sigma [H(B)_\omega - H(B|X)_\omega] && (19.56)\\
&= \max_\sigma [H(B)_\omega - H(E|X)_\omega], && (19.57)
\end{aligned}$$

where the second equality follows from the definition of the quantum mutual information, and the third equality follows because, conditional on X, the input to the channel is pure and the entropies $H(B|X)_\omega$ and $H(E|X)_\omega$ are equal.

Exercise 19.4.1 Prove that it is sufficient to consider pure state inputs when maximizing the following entropy difference over classical-quantum states:

$$\max_\sigma [H(B)_\omega - H(E|X)_\omega]. \tag{19.58}$$

Suppose now that σ is a state that maximizes the Holevo information of the joint channel $\mathcal{N}_{\text{H}} \otimes \mathcal{N}$, and suppose it has the following form:

$$\sigma^{XA'_1A'_2} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes |\phi_x\rangle\langle\phi_x|^{A'_1A'_2}, \tag{19.59}$$

Let

$$\omega^{XB_1B_2E_1E_2} \equiv (U_{\mathcal{N}_{\text{H}}}^{A'_1\to B_1E_1} \otimes U_{\mathcal{N}}^{A'_2\to B_2E_2})(\sigma^{XA'_1A'_2}). \tag{19.60}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

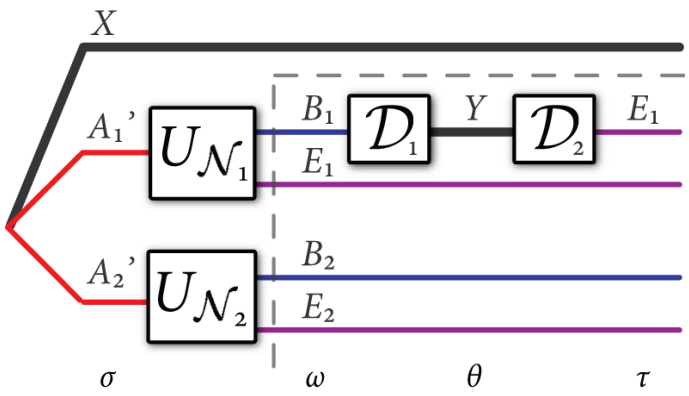

Figure 19.4: A summary of the structural relationships for the additivity question if one channel is a quantum Hadamard channel. Alice first prepares a state of the form in (19.59). She transmits one system A_1' through the quantum Hadamard channel and the other A_2' through the other channel. The first Bob B_1 at the output of the Hadamard channel can simulate the channel to the first Eve E_1 because the first channel is a quantum Hadamard channel. He performs a von Neumann measurement of his system, leading to a classical variable Y, followed by the preparation of some state conditional on the value of the classical variable Y. The bottom of the figure labels the state of the systems at each step.

The Hadamard channel is degradable, and the degrading map from Bob to Eve takes a particular form: it is a von Neumann measurement that produces a classical variable Y, followed by the preparation of a state conditional on the outcome of the measurement. Let $\mathcal{D}_1^{B_1\to Y}$ be the first part of the degrading map that produces the classical variable Y, and let $\theta^{XYE_1B_2E_2} \equiv \mathcal{D}_1^{B_1\to Y}(\omega^{XB_1B_2E_1E_2})$. Let $\mathcal{D}_2^{Y\to E_1}$ be the second part of the degrading channel that produces the state of E_1 conditional on the classical variable Y, and let $\tau^{XE_1E_1B_2E_2} \equiv \mathcal{D}_2^{Y\to E_1}(\theta^{XYE_1B_2E_2})$. Figure 19.4 summarizes these structural relationships. Consider the following chain of inequalities:

$$\begin{aligned}
I(X;B_1B_2)_\omega &= H(B_1B_2)_\omega - H(B_1B_2|X)_\omega && (19.61)\\
&= H(B_1B_2)_\omega - H(E_1E_2|X)_\omega && (19.62)\\
&\leq H(B_1)_\omega + H(B_2)_\omega - H(E_1|X)_\omega - H(E_2|E_1X)_\omega && (19.63)\\
&= H(B_1)_\omega - H(E_1|X)_\omega + H(B_2)_\omega - H(E_2|E_1X)_\tau && (19.64)\\
&\leq H(B_1)_\omega - H(E_1|X)_\omega + H(B_2)_\theta - H(E_2|YX)_\theta && (19.65)\\
&\leq \chi(\mathcal{N}_{\mathrm{H}}) + \chi(\mathcal{N}). && (19.66)
\end{aligned}$$

The first equality follows from the definition of the quantum mutual information. The second equality follows because $H(B_1B_2|X)_\omega = H(E_1E_2|X)_\omega$ when the conditional inputs $|\phi_x\rangle^{A_1'A_2'}$ to the channel are pure states. The next inequality follows from subadditivity of entropy $H(B_1B_2)_\omega \leq H(B_1)_\omega + H(B_2)_\omega$ and from the chain rule for entropy: $H(E_1E_2|X)_\omega = H(E_1|X)_\omega + H(E_2|E_1X)_\omega$. The third equality follows from a rearrangement of terms and realizing that the state of τ on systems E_1E_2X is equivalent to the state of ω on the same systems. The second inequality follows from the quantum data processing

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

inequality $I(E_2; E_1|X)_\tau \leq I(E_2; Y|X)_\theta$. The final inequality follows because the state ω is a state of the form in (19.57), because the entropy difference is never greater than the Holevo information of the first channel, and from the result of Exercise 19.4.1. The same reasoning follows for the other entropy difference and by noting that the classical system is the composite system XY. □

19.4.3 Classical Capacity of the Depolarizing Channel

The qudit depolarizing channel is another example of a channel for which we can compute its classical capacity. Additionally, we will see that achieving the classical capacity of this channel requires a strategy which is very "classical"—it is sufficient to prepare classical states $\{|x\rangle\langle x|\}$ at the input of the channel and to measure each channel output in the same basis (see Exercise 19.4.3). Though, we will later see in Chapter 23 that the depolarizing channel has some rather bizarre, uniquely quantum features when considering its quantum capacity, even though the features of its classical capacity are rather classical.

Recall from Section 4.4.6 that the depolarizing channel is the following map:

$$\mathcal{N}_{\mathrm{D}}(\rho) = (1-p)\rho + p\pi, \tag{19.67}$$

where π is the maximally mixed state.

Theorem 19.4.2 (Classical Capacity of the Depolarizing Channel)**.** *The classical capacity of the qudit depolarizing channel $\mathcal{N}_D$ is as follows:*

$$\chi(\mathcal{N}_D) = \log_2 d + \left(1 - p + \frac{p}{d}\right)\log_2\left(1 - p + \frac{p}{d}\right) + (d-1)\frac{p}{d}\log_2\left(\frac{p}{d}\right), \tag{19.68}$$

Proof. The first part of the proof of this theorem relies on a somewhat technical result, namely, that the Holevo information of the tensor product channel $\mathcal{N}_{\mathrm{D}} \otimes \mathcal{N}$ is additive (where the first channel is the depolarizing channel and the other is arbitrary):

$$\chi(\mathcal{N}_{\mathrm{D}} \otimes \mathcal{N}) = \chi(\mathcal{N}_{\mathrm{D}}) + \chi(\mathcal{N}). \tag{19.69}$$

This result is due to King [171], and it exploits a few properties of the depolarizing channel. The result implies that the classical capacity of the depolarizing channel is equal to its Holevo information. We now show how to compute the Holevo information of the depolarizing channel. To do so, we first determine the minimum output entropy of the channel.

Definition 19.4.1 (Minimum Output Entropy)**.** *The minimum output entropy $H_{\min}(\mathcal{N})$ of a channel $\mathcal{N}$ is the minimum of the entropy at the output of the channel:*

$$H_{\min}(\mathcal{N}) \equiv \min_{\rho} H(\mathcal{N}(\rho)), \tag{19.70}$$

where the minimization is over all states input to the channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 19.4.2 Prove that it is sufficient to minimize over only pure state input states to the channel when computing the minimum output entropy. That is,

$$H_{\min}(\mathcal{N}) = \min_{|\psi\rangle} H(\mathcal{N}(|\psi\rangle\langle\psi|)). \tag{19.71}$$

The depolarizing channel is a highly symmetric channel. For example, if we input a pure state $|\psi\rangle$ to the channel, the output is as follows:

$$\begin{aligned}(1-p)\psi + p\pi &= (1-p)\psi + \frac{p}{d}I && (19.72)\\ &= (1-p)\psi + \frac{p}{d}(\psi + I - \psi) && (19.73)\\ &= \Big(1-p+\frac{p}{d}\Big)\psi + \frac{p}{d}(I-\psi) && (19.74)\end{aligned}$$

Observe that the eigenvalues of the output state are the same for any pure state and are equal to $1-p+\frac{p}{d}$ with multiplicity one and $\frac{p}{d}$ with multiplicity $d-1$. Thus, the minimum output entropy of the depolarizing channel is just

$$H_{\min}(\mathcal{N}_{\mathrm{D}}) = -\Big(1-p+\frac{p}{d}\Big)\log_2\Big(1-p+\frac{p}{d}\Big) - (d-1)\frac{p}{d}\log_2\Big(\frac{p}{d}\Big). \tag{19.75}$$

We now compute the Holevo information of the depolarizing channel. Recall from Theorem 12.3.2 that it is sufficient to consider optimizing the Holevo information over a classical-quantum state with conditional states that are pure (a state $\sigma^{XA'}$ of the form in (19.54)). Also, the Holevo information has the following form:

$$\max_{\sigma} I(X;B)_{\omega} = \max_{\sigma}[H(B)_{\omega} - H(B|X)_{\omega}], \tag{19.76}$$

where ω^{XB} is the output state. Consider the following augmented input ensemble:

$$\rho^{XIJA'} \equiv \frac{1}{d^2}\sum_{x}\sum_{i,j=0}^{d-1} p_X(x)|x\rangle\langle x|^X \otimes |i\rangle\langle i|^I \otimes |j\rangle\langle j|^J \otimes X(i)Z(j)\psi_x^{A'}Z^\dagger(j)X^\dagger(i), \tag{19.77}$$

where $X(i)$ and $Z(j)$ are the generalized Pauli operators from Section 3.6.2. Suppose that we trace over the IJ system. Then the state $\rho^{XA'}$ is as follows:

$$\rho^{XA'} = \sum_x p_X(x)|x\rangle\langle x|^X \otimes \pi^{A'}, \tag{19.78}$$

by recalling the result of Exercise 4.4.9. Also, note that inputting the maximally mixed state to the depolarizing channel results in the maximally mixed state at its output. Consider the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

following chain of inequalities:

$$\begin{aligned} I(X;B)_\omega &= H(B)_\omega - H(B|X)_\omega && (19.79)\\ &\leq H(B)_\rho - H(B|X)_\omega && (19.80)\\ &= \log_2 d - H(B|XIJ)_\rho && (19.81)\\ &= \log_2 d - \sum_x p_X(x) H(B)_{\mathcal{N}_D(\psi_x^{A'})} && (19.82)\\ &\leq \log_2 d - \min_x H(B)_{\mathcal{N}_D(\psi_x^{A'})} && (19.83)\\ &\leq \log_2 d - H_{\min}(\mathcal{N}_{\mathrm{D}}) && (19.84) \end{aligned}$$

The first equality follows by expanding the quantum mutual information. The first inequality follows from concavity of entropy. The second equality follows because the state of ρ on system B is the maximally mixed state π and from the following chain of equalities:

$$\begin{aligned} H(B|XIJ)_\rho &= \frac{1}{d^2}\sum_x \sum_{i,j=0}^{d-1} p_X(x) H(B)_{\mathcal{N}_D(X(i)Z(j)\psi_x^{A'}Z^\dagger(j)X^\dagger(i))} && (19.85)\\ &= \frac{1}{d^2}\sum_x \sum_{i,j=0}^{d-1} p_X(x) H(B)_{X(i)Z(j)\mathcal{N}_D(\psi_x^{A'})Z^\dagger(j)X^\dagger(i)} && (19.86)\\ &= \sum_x p_X(x) H(B)_{\mathcal{N}_D(\psi_x^{A'})} && (19.87)\\ &= H(B|X)_\omega. && (19.88) \end{aligned}$$

The third equality in (19.82) follows from the above chain of equalities. The second inequality in (19.83) follows because the expectation is always more than the minimum (this step is not strictly necessary for the depolarizing channel). The last inequality follows because $\min_x H(B)_{\mathcal{N}_D(\psi_x^{A'})} \geq H_{\min}(\mathcal{N}_{\mathrm{D}})$ (though it is actually an equality for the depolarizing channel). An ensemble of the following form suffices to achieve the classical capacity of the depolarizing channel:

$$\frac{1}{d}\sum_{i=0}^{d-1} |i\rangle\langle i|^I \otimes |i\rangle\langle i|^{A'}, \tag{19.89}$$

because we only require that the reduced state on A' be equivalent to the maximally mixed state. The final expression for the classical capacity of the depolarizing channel is as stated in Theorem 19.4.2, which we plot in Figure 19.5 as a function of the dimension d and the depolarizing parameter p. □

Exercise 19.4.3 (Achieving the classical capacity of the depolarizing channel) We actually know that even more is true regarding the method for achieving the classical capacity of the depolarizing channel. Prove that it is possible to achieve the classical capacity of the depolarizing channel by choosing states from an ensemble $\left\{\frac{1}{d}, |x\rangle\langle x|\right\}$ and performing a von Neumann measurement in the same basis at the output of each channel. That is, the naive

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

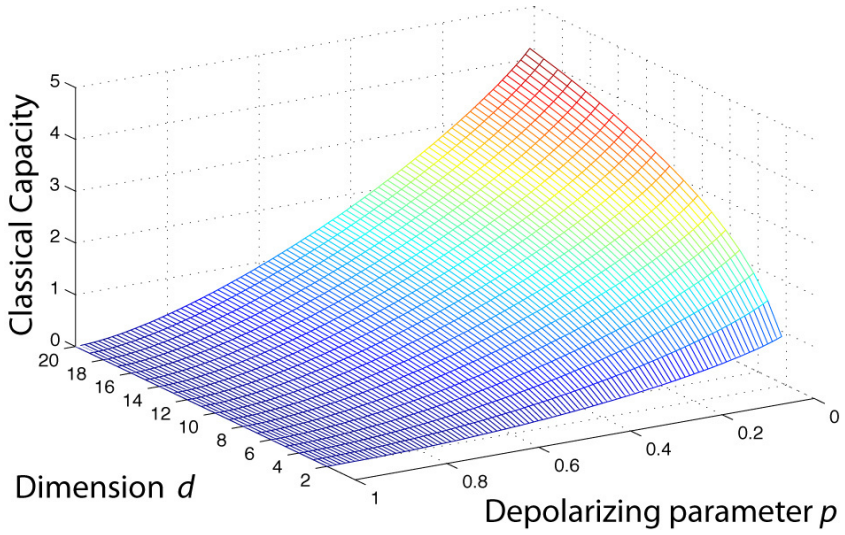

Figure 19.5: The classical capacity of the quantum depolarizing channel as a function of the dimension d of the channel and the depolarizing parameter p. The classical capacity vanishes when $p = 1$ because the channel replaces the input with the maximally mixed state. The classical capacity is maximal at $\log_2 d$ when $p = 0$ because there is no noise. In between these two extremes, the classical capacity is a smooth function of p and d given by the expression in (19.68).

scheme outlined in Section 19.1 is sufficient to attain the classical capacity of the depolarizing channel. (Hint: First show that the classical channel $p_{Y|X}(y|x)$ induced by inputting a state $|x\rangle$ to the depolarizing channel and measuring $|y\rangle$ at the output is as follows:

$$p_{Y|X}(y|x) = (1-p)\delta_{x,y} + \frac{p}{d}. \tag{19.90}$$

Then show that the distribution $p_Y(y)$ is uniform if $p_X(x)$ is uniform. Finally, show that

$$H(Y|X) = -\left(1-p+\frac{p}{d}\right)\log_2\left(1-p+\frac{p}{d}\right) - (d-1)\left(\frac{p}{d}\right)\log_2\left(\frac{p}{d}\right). \tag{19.91}$$

Conclude that the classical capacity of the induced channel $p_{Y|X}(y|x)$ is the same as that for the quantum depolarizing channel.)

Exercise 19.4.4 A covariant channel $\mathcal{N}_{\mathrm{C}}$ is one for which the state resulting from a unitary U acting on the input state before the channel occurs is equivalent to one where there is an irreducible representation of the unitary R_U acting on the output of the channel:

$$\mathcal{N}_{\mathrm{C}}\left(U\rho U^\dagger\right) = R_U \mathcal{N}_{\mathrm{C}}(\rho) R_U^\dagger. \tag{19.92}$$

Show that the Holevo information $\chi(\mathcal{N}_{\mathrm{C}})$ of a covariant channel is equal to

$$\chi(\mathcal{N}_{\mathrm{C}}) = \log_2 d - H(\mathcal{N}_{\mathrm{C}}(\psi)), \tag{19.93}$$

where ψ is an arbitrary pure state.

Exercise 19.4.5 Compute the classical capacity of the quantum erasure channel. First show that it is single-letter. Then show that the classical capacity is equal to $1 - \epsilon$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

19.5 Superadditivity of the Holevo Information

Many researchers thought for some time that the Holevo information would be additive for all quantum channels, implying that it would be a good characterization of the classical capacity in the general case—this conjecture was known as the additivity conjecture. Researchers thought that this conjecture would hold because they discovered a few channels for which it did hold, but without any common theme occurring in the proofs for the different channels, they soon began looking in the other direction for a counterexample to disprove it. After some time, Hastings found the existence of a counterexample to the additivity conjecture, demonstrating that it cannot hold in the general case. This result demonstrates that even one of the most basic questions in quantum Shannon theory still remains wide open and that entanglement at the encoder can help increase classical communication rates over a quantum channel.

We first review a relation between the Holevo information and the minimum output entropy of a tensor product channel. Suppose that we have two channels $\mathcal{N}_1$ and $\mathcal{N}_2$. The Holevo information of the tensor-product channel is additive if

$$\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) = \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2). \tag{19.94}$$

Since the Holevo information is always superadditive for any two channels:

$$\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) \geq \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2), \tag{19.95}$$

(recall the statement at the beginning of the proof of Theorem 12.3.1), we say that it is non-additive if it is strictly superadditive:

$$\chi(\mathcal{N}_1 \otimes \mathcal{N}_2) > \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2). \tag{19.96}$$

The minimum output entropy $H_{\min}(\mathcal{N}_1 \otimes \mathcal{N}_2)$ of the tensor-product channel is a quantity related to the Holevo information (see Definition 19.4.1). It is additive if

$$H_{\min}(\mathcal{N}_1 \otimes \mathcal{N}_2) = H_{\min}(\mathcal{N}_1) + H_{\min}(\mathcal{N}_2). \tag{19.97}$$

Since the minimum output entropy is always subadditive:

$$H_{\min}(\mathcal{N}_1 \otimes \mathcal{N}_2) \leq H_{\min}(\mathcal{N}_1) + H_{\min}(\mathcal{N}_2), \tag{19.98}$$

we say that it is non-additive if it is strictly subadditive:

$$H_{\min}(\mathcal{N}_1 \otimes \mathcal{N}_2) < H_{\min}(\mathcal{N}_1) + H_{\min}(\mathcal{N}_2). \tag{19.99}$$

Additivity of these two quantities is in fact related—it is possible to show that additivity of the Holevo information implies additivity of the minimum output entropy and vice versa (we leave one of these implications as an exercise). Thus, researchers considered additivity of minimum output entropy rather than additivity of Holevo information because it is a simpler quantity to manipulate.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 19.5.1 Prove that non-additivity of the minimum output entropy implies non-additivity of the Holevo information:

$$H_{\min}(\mathcal{N}_1 \otimes \mathcal{N}_2) < H_{\min}(\mathcal{N}_1) + H_{\min}(\mathcal{N}_2) \quad \Rightarrow \quad \chi(\mathcal{N}_1 \otimes \mathcal{N}_2) > \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2). \quad (19.100)$$

(*Hint*: Consider an augmented version $\mathcal{N}_i'$ of each channel $\mathcal{N}_i$, that has its first input be the same as the input to $\mathcal{N}_i$ and its second input be a control input, and the action of the channel is equivalent to measuring the auxiliary input σ and applying a generalized Pauli operator:

$$\mathcal{N}_i'(\rho \otimes \sigma) \equiv \sum_{k,l} X(k)Z(l)\mathcal{N}_i(\rho)Z^\dagger(l)X^\dagger(k)\ \langle k|\langle l|\sigma|k\rangle|l\rangle. \quad (19.101)$$

What is the Holevo information of the augmented channel $\mathcal{N}_i'$? What is the Holevo information of the tensor product of the augmented channels $\mathcal{N}_1' \otimes \mathcal{N}_2'$?) After proving the above statement, we can also conclude that additivity of the Holevo information implies additivity of the minimum output entropy.

We briefly overview the main ideas behind the construction of a channel for which the Holevo information is not additive. Consider a random-unitary channel of the following form:

$$\mathcal{E}(\rho) \equiv \sum_{i=1}^{D} p_i U_i \rho U_i^\dagger, \quad (19.102)$$

where the dimension of the input state is N and the number of random unitaries is D. This channel is "random-unitary" because it applies a particular unitary U_i with probability p_i to the state ρ. The cleverness behind the construction is not actually to provide a deterministic instance of this channel, but rather, to provide a random instance of the channel where both the distribution and the unitaries are chosen at random, and the dimension N and the number D of chosen unitaries satisfy the following relationships:

$$1 \ll D \ll N. \quad (19.103)$$

The other channel to consider to disprove additivity is the conjugate channel

$$\overline{\mathcal{E}}(\rho) \equiv \sum_{i=1}^{D} p_i \overline{U}_i \rho \overline{U}_i^\dagger, \quad (19.104)$$

where p_i and U_i are the same respective probability distribution and unitaries from the channel $\mathcal{E}$, and $\overline{U}_i$ denotes the complex conjugate of U_i. The goal is then to show that there is a non-zero probability over all channels of these forms that the minimum output entropy is non-additive:

$$H_{\min}\big(\mathcal{E} \otimes \overline{\mathcal{E}}\big) < H_{\min}(\mathcal{E}) + H_{\min}\big(\overline{\mathcal{E}}\big). \quad (19.105)$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

A good candidate for a state that could saturate the minimum output entropy $H_{\min}\big(\mathcal{E}\otimes\overline{\mathcal{E}}\big)$ of the tensor-product channel is the maximally entangled state $|\Phi\rangle$, where

$$|\Phi\rangle \equiv \frac{1}{\sqrt{N}}\sum_{i=0}^{N-1}|i\rangle|i\rangle. \tag{19.106}$$

Consider the effect of the tensor-product channel $\mathcal{E}\otimes\overline{\mathcal{E}}$ on the maximally entangled state Φ:

$$\begin{aligned}
&\big(\mathcal{E}\otimes\overline{\mathcal{E}}\big)(\Phi)\\
&=\sum_{i,j=1}^{D} p_i p_j \big(U_i\otimes\overline{U}_j\big)\Phi\Big(U_i^\dagger\otimes\overline{U}_j^\dagger\Big) &(19.107)\\
&=\sum_{i=j} p_i^2 \big(U_i\otimes\overline{U}_i\big)\Phi\Big(U_i^\dagger\otimes\overline{U}_i^\dagger\Big)+\sum_{i\neq j} p_i p_j \big(U_i\otimes\overline{U}_j\big)\Phi\Big(U_i^\dagger\otimes\overline{U}_i^\dagger\Big) &(19.108)\\
&=\left(\sum_{i=1}^{D} p_i^2\right)\Phi+\sum_{i\neq j} p_i p_j \big(U_i\otimes\overline{U}_j\big)\Phi\Big(U_i^\dagger\otimes\overline{U}_i^\dagger\Big), &(19.109)
\end{aligned}$$

where the last line uses the fact that $(M\otimes I)|\Phi\rangle = \big(I\otimes M^T\big)|\Phi\rangle$ for any operator M (this implies that $\big(U\otimes\overline{U}\big)|\Phi\rangle = |\Phi\rangle$). When comparing the above state to one resulting from inputting a product state to the channel, there is a sense in which the above state is less noisy than the product state because D of the combinations of the random unitaries (the ones which have the same index) have no effect on the maximally entangled state. Using techniques from Ref. [125], we can make this intuition precise and obtain the following upper bound on the minimum output entropy:

$$\begin{aligned}
H_{\min}\big(\mathcal{E}\otimes\overline{\mathcal{E}}\big) &\leq H\big(\big(\mathcal{E}\otimes\overline{\mathcal{E}}\big)(\Phi)\big) &(19.110)\\
&\leq 2\ln D-\frac{\ln D}{D}, &(19.111)
\end{aligned}$$

for N and D large enough. Though, using techniques in the same paper, we can also show that

$$H_{\min}(\mathcal{E})\geq \ln D-\delta S^{\max}, \tag{19.112}$$

where

$$\delta S^{\max}\equiv\frac{c}{D}+\text{poly}(D)O\left(\sqrt{\frac{\ln N}{N}}\right), \tag{19.113}$$

c is a constant, and poly(D) indicates a term polynomial in D. Thus, for large enough D and N, it follows that

$$2\delta S^{\max}<\frac{\ln D}{D}, \tag{19.114}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and we get the existence of a channel for which a violation of additivity occurs, because

$$H_{\min}\big(\mathcal{E}\otimes\overline{\mathcal{E}}\big) \leq 2\ln D - \frac{\ln D}{D} \tag{19.115}$$

$$< 2\ln D - 2\delta S^{\max} \tag{19.116}$$

$$\leq H_{\min}(\mathcal{E}) + H_{\min}\big(\overline{\mathcal{E}}\big). \tag{19.117}$$

19.6 Concluding Remarks

The Holevo-Schumacher-Westmoreland (HSW) theorem offers a good characterization of the classical capacity of certain classes of channels, but at the same time, it also demonstrates our lack of understanding of classical transmission over general quantum channels. To be more precise, the Holevo information is a useful characterization of the classical capacity of a quantum channel whenever it is additive, but the regularized Holevo information is not particularly useful as a characterization of it because we cannot even compute this quantity. This suggests that there could be some other formula that better characterizes the classical capacity (if such a formula were additive). As of the writing of this book, such a formula is unknown.

Despite the drawbacks of the HSW theorem, it is still interesting because it at least offers a step beyond the most naive characterization of the classical capacity of a quantum channel with the regularized accessible information. The major insight of HSW was the construction of an explicit POVM (corresponding to a random choice of code) that allows the sender and receiver to communicate at a rate equal to the Holevo information of the channel. This theorem is also useful for determining achievable rates in different communication scenarios: for example, when two senders are trying to communicate over a noisy medium to a single receiver and when a single sender is trying to transmit both classical and quantum information to a receiver.

The depolarizing channel is an example of a quantum channel for which there is a simple expression for its classical capacity. Furthermore, the expression reveals that the scheme needed to achieve the capacity of the channel is rather classical—it is only necessary for the sender to select codewords uniformly at random from some orthonormal basis, and it is only necessary for the receiver to perform measurements of the individual channel outputs in the same orthonormal basis. Thus, the coding scheme is classical because entanglement plays no role at the encoder and the decoding measurements act on the individual channel outputs.

Finally, we discussed Hastings' construction of a quantum channel for which the heralded additivity conjecture does not hold. That is, there exists a channel where entanglement at the encoder can improve communication rates. This superadditive effect is a uniquely quantum phenomenon (recall that Theorem 12.1.1 states that the classical mutual information of a classical channel is additive, and thus correlations at the input cannot increase capacity). This result implies that our best known characterization of the classical capacity of a quantum channel in terms of the channel's Holevo information is far from being a satisfactory characterization of the true capacity, and we still have much more to discover here.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

19.7 History and Further Reading

Holevo was the first to prove the bound bearing his name, regarding the transmission of classical information with a noisy quantum channel [142], and Holevo [144], Schumacher and Westmoreland [219] many years later proved that the Holevo information is an achievable rate for classical data transmission. Just prior to these works, Hausladen *et al.* proved achievability of the Holevo information for the special case of a channel that accepts a classical input and outputs a pure state conditional on the input [126]. They also published a preliminary article [127] in which they answered the catchy question (for the special case of pure states), "How many bits can you fit into a quantum-mechanical it?"

King first proved additivity of the Holevo information for unital qubit channels [170] and later showed it for the depolarizing channel [171]. Shor later showed the equivalence of several additivity conjectures [228] (that they are either all true or all false). Hayden [131], Winter [258], and a joint paper between them [136] proved some results leading up to the work of Hastings [125], who demonstrated a counterexample to the additivity conjecture. Thus, by Shor's aforementioned paper, all of the additivity conjectures are false in general. There has been much follow-up work in an attempt to understand Hastings' result [99, 47, 98, 11].

Some other papers have tried to understand the HSW coding theorem from the perspective of hypothesis testing. Hayashi began much of this work [130], and he covers quite a bit of quantum hypothesis testing in his book [128]. Datta and Mosonyi followed up with some work along these lines [192], as did Renner and Wang [244].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 20

Entanglement-Assisted Classical Communication

We have learned that shared entanglement is often helpful in quantum communication. This is certainly true for the case of a noiseless qubit channel. Without shared entanglement, the most classical information that a sender can reliably transmit over a noiseless qubit channel is just one classical bit (recall Exercise 4.2.2 and the Holevo bound in Exercise 11.9.1). With shared entanglement, they can achieve the super-dense coding resource inequality from Chapter 7:

$$[q \to q] + [qq] \geq 2[c \to c]. \tag{20.1}$$

That is, with one noiseless qubit channel and one shared noiseless ebit, the sender can reliably transmit two classical bits.

A natural question then for us to consider is whether shared entanglement could be helpful in transmitting classical information over a noisy quantum channel $\mathcal{N}$. As a first simplifying assumption, we let Alice and Bob have access to an infinite supply of entanglement, in whatever form they wish, and we would like to know how much classical information Alice can reliably transmit to Bob over such an entanglement-assisted quantum channel. That is, we would like to determine the highest achievable rate C of classical communication in the following resource inequality:

$$\langle \mathcal{N} \rangle + \infty[qq] \geq C[c \to c]. \tag{20.2}$$

The answer to this question is one of the strongest known results in quantum Shannon theory, and it is given by the entanglement-assisted classical capacity theorem. This theorem states that the mutual information $I(\mathcal{N})$ of a quantum channel $\mathcal{N}$ is equal to its entanglement-assisted classical capacity, where

$$I(\mathcal{N}) \equiv \max_{\phi^{AA'}} I(A;B)_\rho, \tag{20.3}$$

$\rho^{AB} \equiv \mathcal{N}^{A' \to B}(\phi^{AA'})$, and the maximization is over all pure bipartite states of the form $\phi^{AA'}$. We should stress that there is no need to regularize this formula in order to characterize the

capacity (as done in the previous chapter and as is so often needed in quantum Shannon theory). The value of this formula *is* the capacity. Also, the optimization task that the formula in (20.3) sets out is a straightforward convex optimization program. Any local maximum is a global maximum because the quantum mutual information is concave in the input state $\phi^{A'}$ (recall Theorem 12.4.2 from Chapter 12) and the set of density operators is convex.

From the perspective of an information theorist, we should only say that a capacity theorem has been solved if there is a tractable formula equal to the optimal rate for achieving a particular operational task. The formula should apply to an arbitrary quantum channel, and it should be a function of that channel. Otherwise, the capacity theorem is still unsolved. There are several operative words in the above sentences that we should explain in more detail. The formula should be tractable, meaning that it sets out an optimization task which is efficient to solve in the dimension of the channel's input system. The formula should give the optimal achievable rate for the given information processing task, meaning that if a rate exceeds the capacity of the channel, then the probability of error for any such protocol should be bounded away from zero as the number of channel uses grows large.[1] Finally, perhaps the most stringent (though related) criterion is that the formula itself (and *not* its regularization) should give the capacity of an arbitrary quantum channel. Despite the success of the HSW coding theorem in demonstrating that the Holevo information of a channel is an achievable rate for classical communication, the classical capacity of a quantum channel is still unsolved because there is an example of a channel for which the Holevo information is not equal to that channel's capacity (see Section 19.5). Thus, it is rather impressive that the formula in (20.3) is equal to the entanglement-assisted classical capacity of an arbitrary channel, given the stringent requirements that we have established for a formula to give the capacity. In this sense, shared entanglement simplifies quantum Shannon theory.

This chapter presents a comprehensive study of the entanglement-assisted classical capacity theorem. We begin by defining the information processing task, consisting of all the steps in a general protocol for classical communication over an entanglement-assisted quantum channel. We then present a simple example of a strategy for entanglement-assisted classical coding that is inspired by dense coding, and in turn, that inspires a strategy for the general case. Section 20.3 states the entanglement-assisted classical capacity theorem. Section 20.4 gives a proof of the direct coding theorem, making use of quantum typicality from Chapter 14, the Packing Lemma from Chapter 15, and ideas in the entanglement concentration protocol from Chapter 18. It demonstrates that the rate in (20.3) is an achievable rate for entanglement-assisted classical communication. After taking a step back from the protocol, we can realize that it is merely a glorified super-dense coding applied to noisy quantum channels. Section 20.5 gives a proof of the converse of the entanglement-assisted classical capacity theorem. It exploits familiar tools such as the Alicki-Fannes' inequality, the quantum data processing inequality, and the chain rule for quantum mutual information

[1] We could strengthen this requirement even more by demanding that the probability of error increases exponentially to one in the asymptotic limit. Fulfilling such a demand would constitute a proof of a *strong converse theorem*.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

(all from Chapter 11), and the last part of it exploits additivity of the mutual information of a quantum channel (from Chapter 12). The converse theorem establishes that the rate in (20.3) is optimal. With the proof of the capacity theorem complete, we then show the interesting result that the classical capacity of a quantum channel assisted by a quantum feedback channel is equal to the entanglement-assisted classical capacity of that channel. We close the chapter by computing the entanglement-assisted classical capacity of both a quantum erasure channel and an amplitude damping channel, and we leave the computation of the entanglement-assisted capacities of two other channels as exercises.

20.1 The Information Processing Task

We begin by explicitly defining the information processing task of entanglement-assisted classical communication, i.e., we define an $(n, C - \delta, \epsilon)$ entanglement-assisted classical code and what it means for a rate C to be achievable. Prior to the start of the protocol, we assume that Alice and Bob share pure-state entanglement in whatever form they wish. For simplicity, we just assume that they share a maximally entangled state of the following form:

$$|\Phi\rangle^{T_A T_B} \equiv \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |i\rangle^{T_A} |i\rangle^{T_B}, \tag{20.4}$$

where the dimension d is as large as they would like it to be. Alice selects some message m uniformly at random from a set $\mathcal{M}$ of messages. Let M denote the random variable corresponding to Alice's random choice of message, and let $|\mathcal{M}|$ be the cardinality of the set $\mathcal{M}$. She applies some CPTP encoding map $\mathcal{E}_m^{T_A \to A'^n}$ to her half of the entangled state $\Phi^{T_A T_B}$ depending on her choice of message m. The global state then becomes

$$\mathcal{E}_m^{T_A \to A'^n}(\Phi^{T_A T_B}). \tag{20.5}$$

Alice transmits the systems A'^n over n independent uses of a noisy channel $\mathcal{N}^{A' \to B}$, leading to the following state

$$\mathcal{N}^{A'^n \to B^n}(\mathcal{E}_m^{T_A \to A'^n}(\Phi^{T_A T_B})), \tag{20.6}$$

where $\mathcal{N}^{A'^n \to B^n} \equiv (\mathcal{N}^{A' \to B})^{\otimes n}$. Bob receives the systems B^n, combines them with his share T_B of the entanglement, and performs a POVM $\{\Lambda_m^{B^n T_B}\}$ on the channel outputs B^n and his share T_B of the entanglement in order to detect the message m that Alice transmits. Figure 20.1 depicts such a general protocol for entanglement-assisted classical communication.

Let M' denote the random variable for the output of Bob's decoding POVM (this represents Bob's estimate of the message). The probability of Bob correctly decoding Alice's message is

$$\Pr\{M' = m \mid M = m\} = \operatorname{Tr}\Big\{\Lambda_m^{B^n T_B} \mathcal{N}^{A'^n \to B^n}(\mathcal{E}_m^{T_A \to A'^n}(\Phi^{T_A T_B}))\Big\}, \tag{20.7}$$

and thus the probability of error $p_e(m)$ for message m is

$$p_e(m) \equiv \operatorname{Tr}\Big\{\big(I - \Lambda_m^{B^n T_B}\big) \mathcal{N}^{A'^n \to B^n}(\mathcal{E}_m^{T_A \to A'^n}(\Phi^{T_A T_B}))\Big\}. \tag{20.8}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

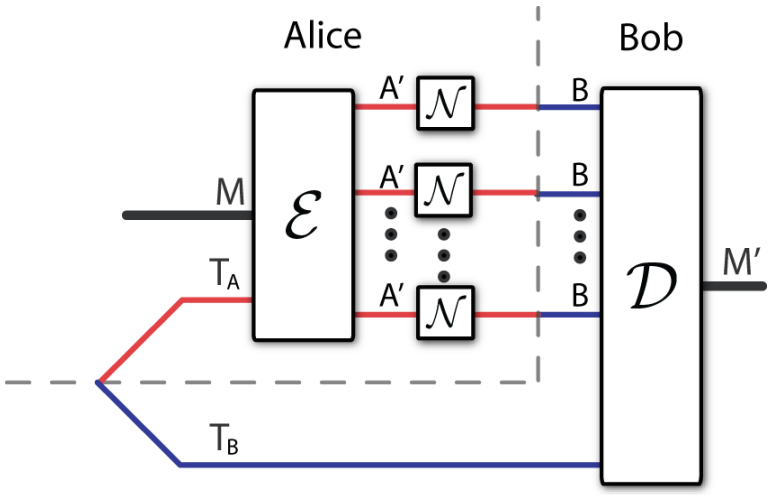

Figure 20.1: The most general protocol for entanglement-assisted classical communication. Alice applies an encoder to her classical message M and her share T_A of the entanglement, and she inputs the encoded systems A'^n to many uses of the channel. Bob receives the outputs of the channel, combines them with his share of the entanglement, and performs some decoding operation to estimate Alice's transmitted message.

The maximal probability of error p_e^* for the coding scheme is

$$p_e^* \equiv \max_{m \in \mathcal{M}} p_e(m). \tag{20.9}$$

The rate C of communication is

$$C \equiv \frac{1}{n} \log_2 |\mathcal{M}| + \delta, \tag{20.10}$$

where δ is an arbitrarily small positive number, and the code has ϵ error if $p_e^* \leq \epsilon$. A rate C of entanglement-assisted classical communication is achievable if there exists an $(n, C - \delta, \epsilon)$ entanglement-assisted classical code for all $\delta, \epsilon > 0$ and sufficiently large n.

20.2 A Preliminary Example

Let us first recall a few items about qudits. The maximally entangled qudit state is

$$|\Phi\rangle^{AB} \equiv \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |i\rangle^A |i\rangle^B. \tag{20.11}$$

Recall from Section 3.6.2 that the Heisenberg-Weyl operators $X(x)$ and $Z(z)$ are an extension of the Pauli matrices to d dimensions:

$$X(x) \equiv \sum_{x'=0}^{d-1} |x + x'\rangle\langle x'|, \qquad Z(z) \equiv \sum_{z'=0}^{d-1} e^{2\pi i z z'/d} |z'\rangle\langle z'|. \tag{20.12}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

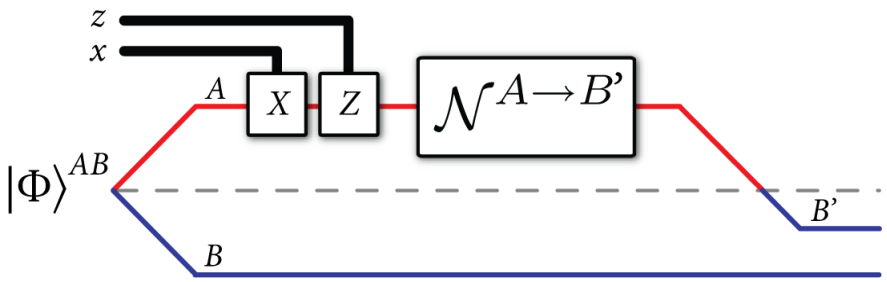

Figure 20.2: A simple scheme, inspired by super-dense coding, for Alice and Bob to exploit shared entanglement and a noisy channel in order to establish an ensemble at Bob's receiving end.

Let $|\Phi_{x,z}\rangle^{AB}$ denote the state that results when Alice applies the operator $X(x)Z(z)$ to her share of the maximally entangled state $|\Phi\rangle^{AB}$:

$$|\Phi_{x,z}\rangle^{AB} \equiv \left(X^A(x)Z^A(z) \otimes I^B\right)|\Phi\rangle^{AB}. \tag{20.13}$$

Recall from Exercise 3.6.11 that the set of states $\left\{|\Phi_{x,z}\rangle^{AB}\right\}_{x,z=0}^{d-1}$ forms a complete orthonormal basis:

$$\langle\Phi_{x',z'}|\Phi_{x,z}\rangle = \delta_{x,x'}\delta_{z,z'}, \qquad \sum_{x,z=0}^{d-1} |\Phi_{x,z}\rangle\langle\Phi_{x,z}| = I^{AB}. \tag{20.14}$$

Let π^{AB} denote the maximally mixed state on Alice and Bob's system: $\pi^{AB} \equiv I^{AB}/d^2$, and let π^A and π^B denote the respective maximally mixed states on Alice and Bob's systems: $\pi^A \equiv I^A/d$ and $\pi^B \equiv I^B/d$. Observe that $\pi^{AB} = \pi^A \otimes \pi^B$.

We now consider a simple strategy, inspired by super-dense coding and the HSW coding scheme from Theorem 19.3.1, that Alice and Bob can employ for entanglement-assisted classical communication. That is, we show how a strategy similar to super-dense coding induces a particular ensemble at Bob's receiving end, to which we can then apply the HSW coding theorem in order to establish the existence of a good code for entanglement-assisted classical communication. Suppose that Alice and Bob possess a maximally entangled qudit state $|\Phi\rangle^{AB}$. Alice chooses two symbols x and z uniformly at random, each in $\{0,\ldots,d-1\}$. She applies the operators $X(x)Z(z)$ to her side of the maximally entangled state $|\Phi\rangle^{AB}$, and the resulting state is $|\Phi_{x,z}\rangle^{AB}$. She then sends her system A over the noisy channel $\mathcal{N}^{A\to B'}$, and Bob receives the output B' from the channel. The noisy channel on the whole system is $\mathcal{N}^{A\to B'} \otimes I^B$, and the ensemble that Bob receives is as follows:

$$\left\{\frac{1}{d^2}, \quad \left(\mathcal{N}^{A\to B'} \otimes I^B\right)\left(\Phi_{x,z}^{AB}\right)\right\}. \tag{20.15}$$

This constitutes an ensemble that they can prepare with one use of the channel and one shared entangled state (Figure 20.2 depicts all of these steps). But, in general, we allow them to exploit many uses of the channel and however much entanglement that they need.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Bob can then perform a collective measurement on both his half of the entanglement and the channel outputs in order to determine a message that Alice is transmitting.

Consider that the above scenario is similar to HSW coding. Theorem 19.3.1 from the previous chapter proves that the Holevo information of the above ensemble is an achievable rate for classical communication over this entanglement-assisted quantum channel. Thus, we can already state and prove the following corollary of Theorem 19.3.1, simply by calculating the Holevo information of the ensemble in (20.15).

Corollary 20.2.1. *The quantum mutual information $I(A;B)_\sigma$ of the state $\sigma^{AB} \equiv \mathcal{N}^{A'\to B}(\Phi^{AA'})$ is an achievable rate for entanglement-assisted classical communication over a quantum channel $\mathcal{N}^{A'\to B}$.*

Proof. Observe that we can map the ensemble in (20.15) to the following classical-quantum state:

$$\rho^{XZB'B} \equiv \sum_{x,z=0}^{d-1} \frac{1}{d^2} |x\rangle\langle x|^X \otimes |z\rangle\langle z|^Z \otimes \left(\mathcal{N}^{A\to B'} \otimes I^B\right)\left(\Phi^{AB}_{x,z}\right). \tag{20.16}$$

The Holevo information of this classical-quantum state is

$$I(XZ;B'B)_\rho = H(B'B)_\rho - H(B'B|XZ)_\rho, \tag{20.17}$$

and it is an achievable rate for entanglement-assisted classical communication over the channel $\mathcal{N}^{A'\to B}$ by Theorem 19.3.1. We now proceed to calculate it. First, we determine the entropy $H(B'B)_\rho$ by tracing over the classical registers XZ:

$$\begin{aligned}
\operatorname{Tr}_{XZ}\left\{\rho^{XZB'B}\right\} &= \sum_{x,z=0}^{d-1} \frac{1}{d^2}\left(\mathcal{N}^{A\to B'} \otimes I^B\right)\left(\Phi^{AB}_{x,z}\right) && (20.18)\\
&= \left(\mathcal{N}^{A\to B'} \otimes I^B\right)\left(\sum_{x,z=0}^{d-1} \frac{1}{d^2}\Phi_{x,z}\right) && (20.19)\\
&= \left(\mathcal{N}^{A\to B'} \otimes I^B\right)\left(\pi^{AB}\right) && (20.20)\\
&= \mathcal{N}^{A\to B'}\left(\pi^A\right) \otimes \pi^B, && (20.21)
\end{aligned}$$

where the third equality follows from (20.14). Thus, the entropy $H(B'B)$ is as follows:

$$H(B'B) = H(\mathcal{N}^{A\to B'}\left(\pi^A\right)) + H\left(\pi^B\right). \tag{20.22}$$

We now determine the conditional quantum entropy $H(B'B|XZ)_\rho$:

$$\begin{aligned}
&H(B'B|XZ)_\rho \\
&= \sum_{x,z=0}^{d-1} \frac{1}{d^2} H\left(\left(\mathcal{N}^{A\to B'} \otimes I^B\right)\left(\Phi^{AB}_{x,z}\right)\right) && (20.23)\\
&= \frac{1}{d^2}\sum_{x,z=0}^{d-1} H\left(\mathcal{N}^{A\to B'}\left[\left(X^A(x)Z^A(z)\right)\left(\Phi^{AB}\right)\left(Z^{\dagger A}(z)X^{\dagger A}(x)\right)\right]\right) && (20.24)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$$= \frac{1}{d^2}\sum_{x,z=0}^{d-1} H\Big(\mathcal{N}^{A\to B'}\Big[\big(Z^T\big)^B(z)\big(X^T\big)^B(x)\big(\Phi^{AB}\big)X^{*B}(x)Z^{*B}(z)\Big]\Big) \tag{20.25}$$

$$= \frac{1}{d^2}\sum_{x,z=0}^{d-1} H\Big(\big(Z^T\big)^B(z)\big(X^T\big)^B(x)\Big[\Big(\mathcal{N}^{A\to B'}\Big)\big(\Phi^{AB}\big)\Big]\big(X^{*B}(x)Z^{*B}(z)\big)\Big) \tag{20.26}$$

$$= H\Big(\mathcal{N}^{A\to B'}\big(\Phi^{AB}\big)\Big) \tag{20.27}$$

The first equality follows because the system XZ is classical (recall the result in Section 11.4.1). The second equality follows from the definition of the state $\Phi^{AB}_{x,z}$. The third equality follows by exploiting the Bell-state matrix identity in Exercise 3.6.12. The fourth equality follows because the unitaries that Alice applies commute with the action of the channel. Finally, the entropy of a state is invariant under any unitaries applied to that state. So the Holevo information $I(XZ;B'B)_\rho$ of the state $\rho^{XZB'B}$ in (20.16) is

$$I(XZ;B'B)_\rho = H\big(\mathcal{N}\big(\pi^A\big)\big) + H\big(\pi^B\big) - H\Big(\Big(\mathcal{N}^{A\to B'}\otimes I^B\Big)\big(\Phi^{AB}\big)\Big), \tag{20.28}$$

Equivalently, we can write it as the following quantum mutual information:

$$I(A;B)_\sigma, \tag{20.29}$$

with respect to the state σ^{AB}, where

$$\sigma^{AB} \equiv \mathcal{N}^{A'\to B}(\Phi^{AA'}). \tag{20.30}$$

□

For some channels, the quantum mutual information in Corollary 20.2.1 is equal to that channel's entanglement-assisted classical capacity. This occurs for the depolarizing channel, a dephasing channel, and an erasure channel to name a few. But there are examples of channels, such as the amplitude damping channel, where the quantum mutual information in Corollary 20.2.1 is not equal to the entanglement-assisted capacity. In the general case, it might perhaps be intuitive that the quantum mutual information of the channel in (20.3) is equal to the entanglement-assisted capacity of the channel, and it is the goal of the next sections to prove this result.

Exercise 20.2.1 Consider the following strategy for transmitting and detecting classical information over an entanglement-assisted depolarizing channel. Alice selects a state $|\Phi_{x,z}\rangle^{AB}$ uniformly at random and sends the A system over the quantum depolarizing channel $\mathcal{N}_D^{A\to B'}$, where

$$\mathcal{N}_D^{A\to B'}(\rho) \equiv (1-p)\rho + p\pi. \tag{20.31}$$

Bob receives the output B' of the channel and combines it with his share B of the entanglement. He then performs a measurement of these systems in the Bell basis $\left\{|\Phi_{x',z'}\rangle\langle\Phi_{x',z'}|^{B'B}\right\}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Determine a simplified expression for the induced classical channel $p_{Z'X'|ZX}(z', x' \mid z, x)$ where

$$p_{Z'X'|ZX}(z', x' \mid z, x) \equiv \langle \Phi_{x',z'}| \left(\mathcal{N}^{A\to B'} \otimes I^B \right) \left(|\Phi_{x,z}\rangle\langle\Phi_{x,z}|^{AB} \right) |\Phi_{x',z'}\rangle. \tag{20.32}$$

Show that the classical capacity of the channel $p_{Z'X'|ZX}(z', x' \mid z, x)$ is equal to the entanglement-assisted classical capacity of the depolarizing channel (you can take it for granted that the entanglement-assisted classical capacity of the depolarizing channel is given by Corollary 20.2.1). Thus, there is no need for the receiver to perform a collective measurement on many channel outputs in order to achieve capacity—it suffices to perform single-channel Bell measurements at the receiver.

20.3 The Entanglement-Assisted Classical Capacity Theorem

We now state the entanglement-assisted classical capacity theorem. Section 20.4 proves the direct part of this theorem, and Section 20.5 proves its converse part.

Theorem 20.3.1 (Bennett-Shor-Smolin-Thapliyal)**.** *The entanglement-assisted classical capacity of a quantum channel is the supremum over all achievable rates for entanglement-assisted classical communication, and it is equal to the channel's mutual information:*

$$\sup\{C \mid C \text{ is achievable}\} = I(\mathcal{N}), \tag{20.33}$$

where the mutual information $I(\mathcal{N})$ of a channel $\mathcal{N}$ is defined as $I(\mathcal{N}) \equiv \max_{\varphi^{AA'}} I(A;B)_\rho$, $\rho^{AB} \equiv \mathcal{N}^{A'\to B}(\varphi^{AA'})$, and $\varphi^{AA'}$ is a pure bipartite state.

20.4 The Direct Coding Theorem

The direct coding theorem is a statement of achievability:

Theorem 20.4.1 (Direct Coding)**.** *The following resource inequality corresponds to an achievable protocol for entanglement-assisted classical communication over a noisy quantum channel:*

$$\langle \mathcal{N} \rangle + H(A)_\rho[qq] \geq I(A;B)_\rho[c \to c], \tag{20.34}$$

where $\rho^{AB} \equiv \mathcal{N}^{A'\to B}(\varphi^{AA'})$.

We suppose that Alice and Bob share n copies of an arbitrary pure, bipartite entangled state $|\varphi\rangle^{AB}$. The amount of entanglement in this state is equivalent to $nH(A)$ ebits. We would like to apply a similar coding technique as outlined in Section 20.2. For example, it would be useful to exploit the transpose trick from Exercise 3.6.12, but we cannot do so directly because this trick only applies to maximally entangled states. Though, we can instead

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

exploit the fact that Alice and Bob share many copies of the state $|\varphi\rangle^{AB}$ that decompose into a direct sum of maximally entangled states. The development is similar to that which we outlined for the entanglement concentration protocol from Chapter 18. First, recall that every pure, bipartite state has a Schmidt decomposition (see Theorem 3.6.1):

$$|\varphi\rangle^{AB} \equiv \sum_x \sqrt{p_X(x)}|x\rangle^A|x\rangle^B, \tag{20.35}$$

where $p_X(x) \geq 0$, $\sum_x p_X(x) = 1$, and $\{|x\rangle^A\}$ and $\{|x\rangle^B\}$ are orthonormal bases for Alice and Bob's respective systems. Let us take n copies of the above state, giving a state of the following form:

$$|\varphi\rangle^{A^nB^n} \equiv \sum_{x^n} \sqrt{p_{X^n}(x^n)}|x^n\rangle^{A^n}|x^n\rangle^{B^n}, \tag{20.36}$$

where

$$x^n \equiv x_1 \cdots x_n, \tag{20.37}$$
$$p_{X^n}(x^n) \equiv p_X(x_1)\cdots p_X(x_n), \tag{20.38}$$
$$|x^n\rangle \equiv |x_1\rangle \cdots |x_n\rangle. \tag{20.39}$$

We can write the above state in terms of its type decomposition (just as we did in (18.29-18.32) for the entanglement concentration protocol):

$$|\varphi\rangle^{A^nB^n} = \sum_t \sum_{x^n \in T_t} \sqrt{p_{X^n}(x^n)}|x^n\rangle^{A^n}|x^n\rangle^{B^n} \tag{20.40}$$
$$= \sum_t \sqrt{p_{X^n}(x_t^n)} \sum_{x^n \in T_t} |x^n\rangle^{A^n}|x^n\rangle^{B^n} \tag{20.41}$$
$$= \sum_t \sqrt{p_{X^n}(x_t^n)d_t}\frac{1}{\sqrt{d_t}} \sum_{x^n \in T_t} |x^n\rangle^{A^n}|x^n\rangle^{B^n} \tag{20.42}$$
$$= \sum_t \sqrt{p(t)}|\Phi_t\rangle^{A^nB^n}, \tag{20.43}$$

with the following definitions:

$$p(t) \equiv p_{X^n}(x_t^n)d_t, \tag{20.44}$$
$$|\Phi_t\rangle^{A^nB^n} \equiv \frac{1}{\sqrt{d_t}} \sum_{x^n \in T_t} |x^n\rangle^{A^n}|x^n\rangle^{B^n}. \tag{20.45}$$

We point the reader to (18.29-18.32) for explanations of these equalities.

Each state $|\Phi_t\rangle^{A^nB^n}$ is maximally entangled with Schmidt rank d_t, and we can thus apply the transpose trick for operators acting on the type class subspaces. Inspired by the dense-coding-like strategy from Section 20.2, we allow Alice to choose unitary operators from the Heisenberg-Weyl set of d_t^2 operators that act on the A^n share of $|\Phi_t\rangle^{A^nB^n}$. We denote one of these operators as $V(x_t, z_t) \equiv X(x_t)Z(z_t)$ where $x_t, z_t \in \{0, \ldots, d_t - 1\}$. If she does this

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

for every type class subspace and applies a phase $(-1)^{b_t}$ in each subspace, then the resulting unitary operator $U(s)$ acting on all of her A^n systems is a direct sum of all of these unitaries:

$$U(s) \equiv \bigoplus_t (-1)^{b_t} V(x_t, z_t), \tag{20.46}$$

where s is a vector containing all of the indices needed to specify the unitary $U(s)$:

$$s \equiv ((x_t, z_t, b_t))_t. \tag{20.47}$$

Let $\mathcal{S}$ denote the set of all possible vectors s. The transpose trick holds for these particular unitary operators:

$$\left(U(s)^{A^n} \otimes I^{B^n}\right)|\varphi\rangle^{A^n B^n} = \left(I^{A^n} \otimes \left(U^T(s)\right)^{B^n}\right)|\varphi\rangle^{A^n B^n} \tag{20.48}$$

because it applies in each type class subspace:

$$\begin{aligned}
&\left(U(s)^{A^n} \otimes I^{B^n}\right)|\varphi\rangle^{A^n B^n} \\
&= \left(\bigoplus_t (-1)^{b_t} V(x_t, z_t)\right)^{A^n} \sum_t \sqrt{p(t)}|\Phi_t\rangle^{A^n B^n} && (20.49)\\
&= \sum_t \sqrt{p(t)}(-1)^{b_t} V(x_t, z_t)^{A^n} |\Phi_t\rangle^{A^n B^n} && (20.50)\\
&= \sum_t \sqrt{p(t)}(-1)^{b_t} V^T(x_t, z_t)^{B^n} |\Phi_t\rangle^{A^n B^n} && (20.51)\\
&= \left(\bigoplus_t (-1)^{b_t} V^T(x_t, z_t)\right)^{B^n} \sum_t \sqrt{p(t)}|\Phi_t\rangle^{A^n B^n} && (20.52)\\
&= \left(I^{A^n} \otimes \left(U^T(s)\right)^{B^n}\right)|\varphi\rangle^{A^n B^n} && (20.53)
\end{aligned}$$

Now we need to establish a means by which Alice can select a random code. For every message $m \in \mathcal{M}$ that Alice would like to transmit, she chooses the elements of the vector $s \in \mathcal{S}$ uniformly at random, leading to a particular unitary operator $U(s)$. We can write $s(m)$ instead of just s to denote the explicit association of the vector s with the message m—we can think of each chosen vector $s(m)$ as a classical codeword, with the codebook being $\{s(m)\}_{m \in \{1,\ldots,|\mathcal{M}|\}}$. This random selection procedure leads to entanglement-assisted quantum codewords of the following form:

$$|\varphi_m\rangle^{A^n B^n} \equiv \left(U(s(m))^{A^n} \otimes I^{B^n}\right)|\varphi\rangle^{A^n B^n}. \tag{20.54}$$

Alice then transmits her systems A^n through many uses of the noisy channel, leading to the following state that is entirely in Bob's control:

$$\mathcal{N}^{A^n \to B'^n}\left(|\varphi_m\rangle\langle\varphi_m|^{A^n B^n}\right). \tag{20.55}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Interestingly, the above state is equal to the state in (20.58) below, by exploiting the transpose trick from (20.48):

$$\begin{aligned}
&\mathcal{N}^{A^n\to B'^n}\left(|\varphi_m\rangle\langle\varphi_m|^{A^nB^n}\right)\\
&=\mathcal{N}^{A^n\to B'^n}\left(U(s(m))^{A^n}|\varphi\rangle\langle\varphi|^{A^nB^n}U^\dagger(s(m))^{A^n}\right) &(20.56)\\
&=\mathcal{N}^{A^n\to B'^n}\left(U^T(s(m))^{B^n}|\varphi\rangle\langle\varphi|^{A^nB^n}U^*(s(m))^{B^n}\right) &(20.57)\\
&=U^T(s(m))^{B^n}\mathcal{N}^{A^n\to B'^n}\left(|\varphi\rangle\langle\varphi|^{A^nB^n}\right)U^*(s(m))^{B^n}. &(20.58)
\end{aligned}$$

Observe that the transpose trick allows us to commute the action of the channel with Alice's encoding unitary $U(s(m))$. Let $\rho^{B'^nB^n}\equiv\mathcal{N}^{A^n\to B'^n}\left(|\varphi\rangle\langle\varphi|^{A^nB^n}\right)$ so that

$$\mathcal{N}^{A^n\to B'^n}\left(|\varphi_m\rangle\langle\varphi_m|^{A^nB^n}\right)=U^T(s(m))^{B^n}\rho^{B'^nB^n}U^*(s(m))^{B^n}. \tag{20.59}$$

Remark 20.4.1 (Tensor-Power Channel Output States) When using the coding scheme given above, the reduced state on the channel output (obtained by ignoring Bob's half of the entanglement in B^n) is a tensor-power state, regardless of the unitary that Alice applies at the channel input:

$$\begin{aligned}
\mathrm{Tr}_{B^n}\left\{\mathcal{N}^{A^n\to B'^n}\left(|\varphi_m\rangle\langle\varphi_m|^{A^nB^n}\right)\right\}&=\rho^{B'^n} &(20.60)\\
&=\mathcal{N}^{A^n\to B'^n}\left(\varphi^{A^n}\right), &(20.61)
\end{aligned}$$

where $\varphi^{A^n}=\left(\mathrm{Tr}_B\left\{\varphi^{AB}\right\}\right)^{\otimes n}$. This follows directly from (20.59) and taking the partial trace over B^n. We exploit this feature in the next chapter, where we construct codes for transmitting both classical and quantum information with the help of shared entanglement.

After Alice has transmitted her entanglement-assisted quantum codewords over the channel, it becomes Bob's task to determine which message m Alice transmitted, and he should do so with some POVM $\{\Lambda_m\}$ that depends on the random choice of code. Figure 20.3 depicts the protocol.

At this point, we would like to exploit the Packing Lemma from Chapter 15 in order to establish the existence of a reliable decoding POVM for Bob. Recall that the Packing Lemma requires four objects, and these four objects should satisfy the four inequalities in (15.11-15.14). The first object required is an ensemble from which Alice and Bob can select a code randomly, and in our case, the ensemble is

$$\left\{\frac{1}{|\mathcal{S}|},\ U^T(s)^{B^n}\rho^{B'^nB^n}U^*(s)^{B^n}\right\}_{s\in\mathcal{S}}. \tag{20.62}$$

The next object required is the expected density operator of this ensemble:

$$\begin{aligned}
\overline{\rho}^{B'^nB^n}&\equiv\mathbb{E}_S\left\{U^T(S)^{B^n}\rho^{B'^nB^n}U^*(S)^{B^n}\right\} &(20.63)\\
&=\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}U^T(s)^{B^n}\rho^{B'^nB^n}U^*(s)^{B^n}. &(20.64)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

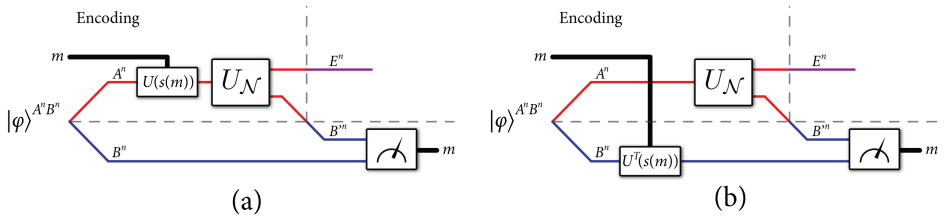

Figure 20.3: (a) Alice shares many copies of a pure, bipartite state $|\varphi\rangle^{\otimes n}$ with Bob. She encodes a message m according to some unitary of the form in (20.46). She transmits her half of the entanglement-assisted quantum codeword over many uses of the quantum channel, and it is Bob's task to determine which message she transmits. (b) Alice acting locally with the unitary $U(s(m))$ on her half A^n of the entanglement $|\varphi\rangle^{\otimes n}$ is the same as her acting nonlocally with $U^T(s(m))$ on Bob's half B^n of the entanglement. This follows because of the particular structure of the unitaries in (20.46).

We later prove that this expected density operator has the following simpler form:

$$\overline{\rho}^{B'^n B^n} = \sum_t p(t)\, \mathcal{N}^{A^n \to B'^n}\left(\pi_t^{A^n}\right) \otimes \pi_t^{B^n}, \tag{20.65}$$

where $p(t)$ is the distribution from (20.44) and π_t is the maximally mixed state on a type class subspace: $\pi_t \equiv I_t \,/\, d_t$. The final two objects that we require for the Packing Lemma are the message subspace projectors and the total subspace projector. We assign these respectively as

$$U^T(s)^{B^n} \Pi_{\rho,\delta}^{B'^n B^n} U^*(s)^{B^n}, \tag{20.66}$$

$$\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}, \tag{20.67}$$

where $\Pi_{\rho,\delta}^{B'^n B^n}$, $\Pi_{\rho,\delta}^{B'^n}$, and $\Pi_{\rho,\delta}^{B^n}$ are the typical projectors for many copies of the states $\rho^{B'B} \equiv \mathcal{N}^{A\to B'}\left(\varphi^{AB}\right)$, $\rho^{B'} = \mathrm{Tr}_B\{\rho^{B'B}\}$, and $\rho^B = \mathrm{Tr}_{B'}\{\rho^{B'B}\}$, respectively. Observe that the size of each message subspace projector is $\approx 2^{nH(B'B)}$, and the size of the total subspace projector is $\approx 2^{n[H(B')+H(B)]}$. By dimension counting, this is suggesting that we can pack in $\approx 2^{n[H(B')+H(B)]} \,/\, 2^{nH(B'B)} = 2^{nI(B';B)}$ messages with this coding technique.

If the four conditions of the Packing Lemma are satisfied (see (15.11-15.14)), then there exists a detection POVM that can reliably decode Alice's transmitted messages as long as the number of messages in the code is not too high. The four conditions in (15.11-15.14) translate to the following four conditions for our case:

$$\mathrm{Tr}\left\{\left(\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\right)\left(U^T(s)^{B^n} \rho^{B'^n B^n} U^*(s)^{B^n}\right)\right\} \geq 1-\epsilon, \tag{20.68}$$

$$\mathrm{Tr}\left\{\left(U^T(s)^{B^n} \Pi_{\rho,\delta}^{B'^n B^n} U^*(s)^{B^n}\right)\left(U^T(s)^{B^n} \rho^{B'^n B^n} U^*(s)^{B^n}\right)\right\} \geq 1-\epsilon, \tag{20.69}$$

$$\mathrm{Tr}\left\{U^T(s)^{B^n} \Pi_{\rho,\delta}^{B'^n B^n} U^*(s)^{B^n}\right\} \leq 2^{n\left[H(B'B)_\rho + c\delta\right]}, \tag{20.70}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$$\left(\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\right) \overline{\rho}^{B'^n B^n} \left(\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\right) \leq 2^{-n\left[H(B')_\rho + H(B)_\rho - \eta(n,\delta) - c\delta\right]} \left(\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\right), \quad (20.71)$$

where c is some positive constant and $\eta(n,\delta)$ is a function that approaches zero as $n \to \infty$ and $\delta \to 0$.

We now prove the four inequalities in (20.68-20.71), attacking them in the order of increasing difficulty. The condition in (20.69) holds because

$$\begin{aligned} &\text{Tr}\left\{ \left(U^T(s)^{B^n} \Pi_{\rho,\delta}^{B'^n B^n} U^*(s)^{B^n}\right) \left(U^T(s)^{B^n} \rho^{B'^n B^n} U^*(s)^{B^n}\right) \right\} \\ &= \text{Tr}\left\{ \Pi_{\rho,\delta}^{B'^n B^n} \rho^{B'^n B^n} \right\} \quad (20.72) \\ &\geq 1 - \epsilon. \quad (20.73) \end{aligned}$$

The equality holds by cyclicity of the trace and because $U^* U^T = I$. The inequality holds by exploiting the unit probability property of typical projectors (Property 14.1.1). From this inequality, observe that we choose each message subspace projector so that it is exactly the one that should identify the entanglement-assisted quantum codeword $U^T(s)^{B^n} \rho^{B'^n B^n} U^*(s)^{B^n}$ with high probability.

We next consider the condition in (20.70):

$$\begin{aligned} \text{Tr}\left\{ U^T(s)^{B^n} \Pi_{\rho,\delta}^{B'^n B^n} U^*(s)^{B^n} \right\} &= \text{Tr}\left\{ \Pi_{\rho,\delta}^{B'^n B^n} \right\} \quad (20.74) \\ &\leq 2^{n\left[H(B'B)_\rho + c\delta\right]}. \quad (20.75) \end{aligned}$$

The equality holds again by cyclicity of trace, and the inequality follows from the exponentially small cardinality property of the typical subspace (Property 14.1.2).

Consider the condition in (20.68). First, define $\hat{P} = I - P$. Then

$$\begin{aligned} \Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n} &= \left(I - \hat{\Pi}_{\rho,\delta}^{B'^n}\right) \otimes \left(I - \hat{\Pi}_{\rho,\delta}^{B^n}\right) \quad (20.76) \\ &= \left(I^{B'^n} \otimes I^{B^n}\right) - \left(\hat{\Pi}_{\rho,\delta}^{B'^n} \otimes I^{B^n}\right) \\ &\quad - \left(I^{B'^n} \otimes \hat{\Pi}_{\rho,\delta}^{B^n}\right) + \left(\hat{\Pi}_{\rho,\delta}^{B'^n} \otimes \hat{\Pi}_{\rho,\delta}^{B^n}\right) \quad (20.77) \\ &\geq \left(I^{B'^n} \otimes I^{B^n}\right) - \left(\hat{\Pi}_{\rho,\delta}^{B'^n} \otimes I^{B^n}\right) - \left(I^{B'^n} \otimes \hat{\Pi}_{\rho,\delta}^{B^n}\right). \quad (20.78) \end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Consider the following chain of inequalities:

$$\begin{aligned}
&\text{Tr}\Big\{\Big(\Pi_{\rho,\delta}^{B'^n}\otimes\Pi_{\rho,\delta}^{B^n}\Big)\Big(U^T(s)^{B^n}\rho^{B'^nB^n}U^*(s)^{B^n}\Big)\Big\}\\
&\geq \text{Tr}\Big\{U^T(s)^{B^n}\rho^{B'^nB^n}U^*(s)^{B^n}\Big\}\\
&\qquad-\text{Tr}\Big\{\Big(\hat{\Pi}_{\rho,\delta}^{B'^n}\otimes I^{B^n}\Big)\Big(U^T(s)^{B^n}\rho^{B'^nB^n}U^*(s)^{B^n}\Big)\Big\}\\
&\qquad\quad-\text{Tr}\Big\{\Big(I^{B'^n}\otimes\hat{\Pi}_{\rho,\delta}^{B^n}\Big)\Big(U^T(s)^{B^n}\rho^{B'^nB^n}U^*(s)^{B^n}\Big)\Big\} &(20.79)\\
&= 1-\text{Tr}\Big\{\hat{\Pi}_{\rho,\delta}^{B'^n}\rho^{B'^n}\Big\}-\text{Tr}\Big\{\hat{\Pi}_{\rho,\delta}^{B^n}\rho^{B^n}\Big\} &(20.80)\\
&\geq 1-2\epsilon. &(20.81)
\end{aligned}$$

The first inequality follows from the development in (20.76-20.78). The first equality follows because $\text{Tr}\Big\{U^T(s)^{B^n}\rho^{B'^nB^n}U^*(s)^{B^n}\Big\}=1$ and from performing a partial trace on B^n and B'^n, respectively (while noting that we can apply the transpose trick for the second one). The final inequality follows from the unit probability property of the typical projectors $\Pi_{\rho,\delta}^{B'^n}$ and $\Pi_{\rho,\delta}^{B^n}$ (Property 14.1.1).

The last inequality in (20.71) requires the most effort to prove. We first need to prove that the expected density operator $\overline{\rho}^{B'^nB^n}$ takes the form given in (20.65). To simplify the development, we evaluate the expectation without the channel applied, and we then apply the channel to the state at the end of the development. Consider that

$$\begin{aligned}
\overline{\rho}^{A^nB^n} &= \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}U^T(s)^{B^n}|\varphi\rangle\langle\varphi|^{A^nB^n}U^*(s)^{B^n} &(20.82)\\
&= \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}U^T(s)^{B^n}\left(\sum_t\sqrt{p(t)}|\Phi_t\rangle^{A^nB^n}\right)\left(\sum_{t'}\langle\Phi_{t'}|^{A^nB^n}\sqrt{p(t')}\right)U^*(s)^{B^n} &(20.83)\\
&= \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\left(\sum_t\sqrt{p(t)}(-1)^{b_t(s)}\big(V^T((z_t,x_t)(s))\big)^{B^n}|\Phi_t\rangle^{A^nB^n}\right)\\
&\qquad\left(\sum_{t'}\langle\Phi_{t'}|^{A^nB^n}(-1)^{b_{t'}(s)}(V^*((z_{t'},x_{t'})(s)))^{B^n}\sqrt{p(t')}\right) &(20.84)
\end{aligned}$$

Let us first consider the case when $t=t'$. Then the expression in (20.84) becomes

$$\begin{aligned}
&\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\sum_t p(t)\big(V^T((z_t,x_t)(s))\big)^{B^n}|\Phi_t\rangle\langle\Phi_t|^{A^nB^n}(V^*((z_t,x_t)(s)))^{B^n} &(20.85)\\
&=\sum_t p(t)\left[\frac{1}{d_t^2}\sum_{x_t,z_t}\big(V^T(z_t,x_t)\big)^{B^n}|\Phi_t\rangle\langle\Phi_t|^{A^nB^n}(V^*(z_t,x_t))^{B^n}\right] &(20.86)\\
&=\sum_t p(t)\pi_t^{A^n}\otimes\pi_t^{B^n}. &(20.87)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

These equalities hold because the sum over all the elements in $\mathcal{S}$ implies that we are uniformly mixing the maximally entangled states $|\Phi_t\rangle^{A^nB^n}$ on the type class subspaces and Exercise 4.4.9 gives us that the resulting state on each type class subspace is equal to $\mathrm{Tr}_{B^n}\{\Phi_t^{A^nB^n}\} \otimes \pi_t^{B^n} = \pi_t^{A^n} \otimes \pi_t^{B^n}$. Let us now consider the case when $t \neq t'$. Then the expression in (20.84) becomes

$$\begin{aligned}
&\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\sum_{t',\ t\neq t'}\sqrt{p(t)p(t')}(-1)^{b_t(s)+b_{t'}(s)}\times \\
&\qquad \left(V^T((z_t,x_t)(s))\right)^{B^n}|\Phi_t\rangle\langle\Phi_{t'}|^{A^nB^n}(V^*((z_{t'},x_{t'})(s)))^{B^n} \\
&= \sum_{t',\ t\neq t'}\frac{1}{d_t^2 d_{t'}^2 4}\sum_{b_t,b_{t'},x_t,z_t,x_{t'},z_{t'}}\sqrt{p(t)p(t')}(-1)^{b_t+b_{t'}}\times \\
&\qquad \left(V^T(z_t,x_t)\right)^{B^n}|\Phi_t\rangle\langle\Phi_{t'}|^{A^nB^n}(V^*(z_{t'},x_{t'}))^{B^n} && (20.88)\\
&= \sum_{t',\ t\neq t'}\frac{1}{d_t^2 d_{t'}^2}\sum_{b_t,b_{t'}}\frac{(-1)^{b_t+b_{t'}}}{4}\times \\
&\qquad \left(\sum_{x_t,z_t,x_{t'},z_{t'}}\sqrt{p(t)p(t')}\left(V^T(z_t,x_t)\right)^{B^n}|\Phi_t\rangle\langle\Phi_{t'}|^{A^nB^n}(V^*(z_{t'},x_{t'}))^{B^n}\right) && (20.89)\\
&= 0 && (20.90)
\end{aligned}$$

It then follows that

$$\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}U^T(s)^{B^n}|\varphi\rangle\langle\varphi|^{A^nB^n}U^*(s)^{B^n} = \sum_t p(t)\pi_t^{A^n}\otimes\pi_t^{B^n}, \tag{20.91}$$

and by linearity, that

$$\begin{aligned}
&\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}U^T(s)^{B^n}\mathcal{N}^{A^n\to B'^n}\left(|\varphi\rangle\langle\varphi|^{A^nB^n}\right)U^*(s)^{B^n} \\
&\qquad\qquad = \sum_t p(t)\mathcal{N}^{A^n\to B'^n}\left(\pi_t^{A^n}\right)\otimes\pi_t^{B^n}. \qquad (20.92)
\end{aligned}$$

We now prove the final condition in (20.71) for the Packing Lemma. Consider the fol-

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

lowing chain of inequalities:

$$\begin{aligned}
&\left(\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\right)\overline{\rho}^{B'^n B^n}\left(\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\right) \\
&= \left(\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\right)\left(\sum_t p(t)\mathcal{N}^{A^n\to B'^n}\left(\pi_t^{A^n}\right) \otimes \pi_t^{B^n}\right)\left(\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\right) && (20.93)\\
&= \sum_t p(t)\left(\Pi_{\rho,\delta}^{B'^n}\mathcal{N}^{A^n\to B'^n}\left(\pi_t^{A^n}\right)\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\pi_t^{B^n}\Pi_{\rho,\delta}^{B^n}\right) && (20.94)\\
&= \sum_t p(t)\left(\Pi_{\rho,\delta}^{B'^n}\mathcal{N}^{A^n\to B'^n}\left(\pi_t^{A^n}\right)\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n}\frac{\Pi_t^{B^n}}{\mathrm{Tr}\{\Pi_t^{B^n}\}}\Pi_{\rho,\delta}^{B^n}\right) && (20.95)\\
&\leq \sum_t p(t)\left(\Pi_{\rho,\delta}^{B'^n}\mathcal{N}^{A^n\to B'^n}\left(\pi_t^{A^n}\right)\Pi_{\rho,\delta}^{B'^n} \otimes 2^{-n\left[H(B)_\rho-\eta(n,\delta)\right]}\Pi_{\rho,\delta}^{B^n}\right) && (20.96)
\end{aligned}$$

The first equality follows from (20.92). The second equality follows by a simple manipulation. The third equality follows because the maximally mixed state $\pi_t^{B^n}$ is equivalent to the normalized type class projection operator $\Pi_t^{X^n}$. The inequality follows from Property 14.3.2 and $\Pi_{\rho,\delta}^{B^n}\Pi_t^{B^n}\Pi_{\rho,\delta}^{B^n} \leq \Pi_{\rho,\delta}^{B^n}$ (the support of a typical type projector is always in the support of the typical projector and the intersection of the support of an atypical type with the typical projector is null). Continuing, by linearity, the last line above is equal to

$$\begin{aligned}
&\Pi_{\rho,\delta}^{B'^n}\mathcal{N}^{A^n\to B'^n}\left(\sum_t p(t)\pi_t^{A^n}\right)\Pi_{\rho,\delta}^{B'^n} \otimes 2^{-n\left[H(B)_\rho-\eta(n,\delta)\right]}\Pi_{\rho,\delta}^{B^n} \\
&= \Pi_{\rho,\delta}^{B'^n}\mathcal{N}^{A^n\to B'^n}\left(\varphi^{A^n}\right)\Pi_{\rho,\delta}^{B'^n} \otimes 2^{-n\left[H(B)_\rho-\eta(n,\delta)\right]}\Pi_{\rho,\delta}^{B^n} && (20.97)\\
&\leq 2^{-n\left[H(B')_\rho-c\delta\right]}\Pi_{\rho,\delta}^{B'^n} \otimes 2^{-n\left[H(B)_\rho-\eta(n,\delta)\right]}\Pi_{\rho,\delta}^{B^n} && (20.98)\\
&= 2^{-n\left[H(B')_\rho+H(B)_\rho-\eta(n,\delta)-c\delta\right]}\Pi_{\rho,\delta}^{B'^n} \otimes \Pi_{\rho,\delta}^{B^n} && (20.99)
\end{aligned}$$

The first equality follows because $\varphi^{A^n} = \sum_t p(t)\pi_t^{A^n}$. The inequality follows from the equipartition property of typical projectors (Property 14.1.3). The final equality follows by rearranging terms.

With the four conditions in (20.68-20.71) holding, it follows from Corollary 15.5.1 (the derandomized version of the Packing Lemma) that there exists a deterministic code and a POVM $\left\{\Lambda_m^{B'^n B^n}\right\}$ that can detect the transmitted states with arbitrarily low maximal probability of error as long as the size $|\mathcal{M}|$ of the message set is small enough:

$$\begin{aligned}
p_e^* &\equiv \max_m \mathrm{Tr}\left\{\left(I-\Lambda_m^{B'^n B^n}\right)U^T(s(m))^{B^n}\rho^{B'^n B^n}U^*(s(m))^{B^n}\right\} && (20.100)\\
&\leq 4\left(\epsilon+2\sqrt{\epsilon}\right) + 8\cdot 2^{-n\left[H(B')_\rho+H(B)_\rho-\eta(n,\delta)-c\delta\right]}2^{n\left[H(B'B)_\rho+c\delta\right]}|\mathcal{M}| && (20.101)\\
&= 4\left(\epsilon+2\sqrt{\epsilon}\right) + 8\cdot 2^{-n\left[I(B';B)_\rho-\eta(n,\delta)-2c\delta\right]}|\mathcal{M}|. && (20.102)
\end{aligned}$$

We can choose the size of the message set to be $|\mathcal{M}| = 2^{n[I(B';B)-\eta(n,\delta)-3c\delta]}$ so that the rate of communication is

$$C = \frac{1}{n}\log_2|\mathcal{M}| = I(B';B) - \eta(n,\delta) - 3c\delta, \tag{20.103}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and the bound on the maximal probability of error becomes

$$p_e^* \leq 4\big(\epsilon + 2\sqrt{\epsilon}\big) + 8 \cdot 2^{-nc\delta}. \tag{20.104}$$

Since ϵ is an arbitrary positive number that approaches zero for sufficiently large n and δ is a positive constant, the maximal probability of error vanishes as n becomes large. Thus, the quantum mutual information $I(B';B)_\rho$, with respect to the state

$$\rho^{B'B} \equiv \mathcal{N}^{A\to B'}\big(\varphi^{AB}\big), \tag{20.105}$$

is an achievable rate for the entanglement-assisted transmission of classical information over $\mathcal{N}$. To obtain the precise statement in Theorem 20.3.1, we can simply rewrite the quantum mutual information as $I(A;B)_\rho$ with respect to the state

$$\rho^{AB} \equiv \mathcal{N}^{A'\to B}\Big(\varphi^{AA'}\Big). \tag{20.106}$$

Alice and Bob can achieve the maximum rate of communication simply by determining the state $\varphi^{AA'}$ that maximizes the quantum mutual information $I(A;B)_\rho$ and by generating entanglement-assisted classical codes from the state ρ^{AB}.

20.5 The Converse Theorem

This section contains the proof of the converse part of the entanglement-assisted classical capacity theorem. Let us begin by supposing that Alice and Bob are trying to use the entanglement-assisted channel many times to accomplish the task of common randomness generation (recall that we took this approach for the converse of the classical capacity theorem in Section 19.3.2).[2] An upper bound on the rate at which Alice and Bob can generate common randomness also serves as an upper bound on the rate at which they can communicate because a noiseless classical channel can generate common randomness. In such a task, Alice and Bob share entanglement in some pure state $|\Phi\rangle^{T_AT_B}$ (though our proof below applies to any shared state). Alice first prepares the maximally correlated state $\overline{\Phi}^{MM'}$, and the rate of common randomness in this state is $C \equiv \frac{1}{n}\log|M|$. Alice then applies some encoding map $\mathcal{E}^{M'T_A\to A^n}$ to the classical system M' and her half T_A of the shared entanglement. The resulting state is

$$\mathcal{E}^{M'T_A\to A^n}(\overline{\Phi}^{MM'} \otimes \Phi^{T_AT_B}). \tag{20.107}$$

She sends her A^n systems through many uses $\mathcal{N}^{A^n\to B^n}$ of the channel $\mathcal{N}^{A\to B}$, and Bob receives the systems B^n, producing the state:

$$\omega^{MT_BB^n} \equiv \mathcal{N}^{A^n\to B^n}(\mathcal{E}^{M'T_A\to A^n}(\overline{\Phi}^{MM'} \otimes \Phi^{T_AT_B})). \tag{20.108}$$

[2] We should qualify in this approach that we are implicitly assuming a bound on the amount of entanglement that they consume in this protocol. Otherwise, they could generate an infinite amount of common randomness. Also, the converse proof outlined here applies equally well if Alice chooses messages from a uniform random variable and tries to communicate this message to Bob.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Finally, Bob performs some decoding map $\mathcal{D}^{B^n T_B \to \hat{M}}$ on the above state to give

$$(\omega')^{M\hat{M}} \equiv \mathcal{D}^{B^n T_B \to \hat{M}}(\omega^{MT_B B^n}). \tag{20.109}$$

If the protocol is ϵ-good for common randomness generation, then the actual state $(\omega')^{M\hat{M}}$ resulting from the protocol should be ϵ-close in trace distance to the ideal common randomness state:

$$\left\| (\omega')^{M\hat{M}} - \overline{\Phi}^{M\hat{M}} \right\|_1 \leq \epsilon. \tag{20.110}$$

We now show that the quantum mutual information of the channel serves as an upper bound on the rate C of any reliable protocol for entanglement-assisted common randomness generation (a protocol meeting the error criterion in (20.110)). Consider the following chain of inequalities:

$$\begin{aligned}
nC &= I(M;\hat{M})_{\overline{\Phi}} && (20.111)\\
&\leq I(M;\hat{M})_{\omega'} + n\epsilon' && (20.112)\\
&\leq I(M;B^n T_B)_{\omega} + n\epsilon' && (20.113)\\
&= I(T_B M;B^n)_{\omega} + I(M;T_B)_{\omega} - I(B^n;T_B)_{\omega} + n\epsilon' && (20.114)\\
&= I(T_B M;B^n)_{\omega} - I(B^n;T_B)_{\omega} + n\epsilon' && (20.115)\\
&\leq I(T_B M;B^n)_{\omega} + n\epsilon' && (20.116)\\
&\leq \max_{\rho^{XAA'^n}} I(AX;B^n)_{\rho} + n\epsilon' && (20.117)
\end{aligned}$$

The first equality follows by evaluating the quantum mutual information of the common randomness state $\overline{\Phi}^{M\hat{M}}$. The first inequality follows from the assumption that the protocol satisfies the error criterion in (20.110) and by applying the Alicki-Fannes' inequality from Exercise 11.9.7 with $\epsilon' \equiv 6\epsilon C + 4H_2(\epsilon)/n$. The second inequality follows from quantum data processing (Corollary 11.9.4)—Bob processes the state ω with the decoder $\mathcal{D}$ to get the state ω'. The second equality follows from the chain rule for quantum mutual information (see Exercise 11.7.1). The third equality follows because the systems M and T_B are in a product state, so $I(M;T_B)_{\omega} = 0$. The third inequality follows because $I(B^n;T_B)_{\omega} \geq 0$. Observe that the state $\omega^{MT_B B^n}$ is a classical-quantum state of the form:

$$\rho^{XAB^n} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A'^n \to B^n}(\rho_x^{AA'^n}), \tag{20.118}$$

where the classical system X in ρ^{XAB^n} plays the role of M in $\omega^{MT_B B^n}$ and the quantum system A in ρ^{XAB^n} plays the role of T_B in $\omega^{MT_B B^n}$. Then the final inequality follows because the quantum mutual information $I(T_B M;B^n)_{\omega}$ can never be greater than the maximum of $I(AX;B^n)_{\rho}$ over all input states of the form in (20.118).

We can strengthen this converse proof considerably. First, observe that the most general form of an encoding is an arbitrary CPTP map $\mathcal{E}^{M'T_A \to A^n}$ that acts on a classical register M'

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and a quantum register T_A. From Section 4.4.8, we know that this map takes the following form:

$$\mathcal{E}^{M'T_A\to A^n}(\overline{\Phi}^{MM'}\otimes\Phi^{T_AT_B})=\frac{1}{M}\sum_m|m\rangle\langle m|^M\otimes\mathcal{E}_m^{T_A\to A^n}(\Phi^{T_AT_B}),\tag{20.119}$$

where each $\mathcal{E}_m^{T_A\to A^n}$ is a CPTP map. This particular form follows because the first register M' on which the map $\mathcal{E}^{M'T_A\to A^n}$ acts is a classical register. Now, it would seem strange if performing a conditional noisy encoding $\mathcal{E}_m^{T_A\to A^n}$ for each message m could somehow improve performance. So, we would like to prove that conditional noisy encodings can never outperform conditional isometric (noiseless) encodings. In this vein, since Alice is in control of the encoder, we allow her to simulate the noisy encodings $\mathcal{E}_m^{T_A\to A^n}$ by acting with their isometric extensions $U_{\mathcal{E}_m}^{T_A\to A^nE'}$ and tracing out the environments E' (to which she has access). Then the value of the quantum mutual information $I(T_BM;B^n)_\omega$ is unchanged by this simulation. Now suppose instead that Alice performs a von Neumann measurement of the environment of the encoding and she places the outcome of the measurement in some classical register L. Then the quantum mutual information can only increase, a result that follows from the quantum data processing inequality:

$$I(T_BLM;B^n)_\omega\geq I(T_BM;B^n)_\omega.\tag{20.120}$$

Thus, isometric encodings are sufficient for achieving the entanglement-assisted classical capacity.

We can view this result in a less operational (and more purely mathematical) way as well. Consider a state of the form in (20.118). Suppose that each $\rho_x^{AA'^n}$ has a spectral decomposition

$$\rho_x^{AA'^n}=\sum_y p_{Y|X}(y|x)\psi_{x,y}^{AA'^n},\tag{20.121}$$

where the states $\psi_{x,y}^{AA'^n}$ are pure. We can define the following augmented state

$$\rho^{XYAB^n}\equiv\sum_{x,y}p_X(x)p_{Y|X}(y|x)|x\rangle\langle x|^X\otimes|y\rangle\langle y|^Y\otimes\mathcal{N}^{A'^n\to B^n}(\psi_{x,y}^{AA'^n}),\tag{20.122}$$

such that $\rho^{XAB^n}=\mathrm{Tr}_Y\{\rho^{XYAB^n}\}$. Then the quantum data processing inequality implies that

$$I(AX;B^n)_\rho\leq I(AXY;B^n)_\rho.\tag{20.123}$$

By joining the classical Y register with the classical X register, the following equality holds

$$\max_{\rho^{XAA'^n}}I(AX;B^n)_\rho=\max_{\sigma^{XAA'^n}}I(AX;B^n)_\sigma,\tag{20.124}$$

where

$$\sigma^{XAB^n}\equiv\sum_x p_X(x)|x\rangle\langle x|^X\otimes\mathcal{N}^{A'^n\to B^n}(\psi_x^{AA'^n}),\tag{20.125}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

so that the maximization is over only pure states $\psi_x^{AA'^n}$. Then we know from the result of Exercise 12.4.1 that

$$\max_{\sigma^{XAA'^n}} I(AX;B^n)_\omega = \max_{\phi^{AA'^n}} I(A;B^n)_\omega, \tag{20.126}$$

where the maximization on the RHS is over pure states $\phi^{AA'^n}$. Finally, from additivity of the quantum mutual information of a quantum channel (Theorem 12.4.1) and an inductive argument similar to that in Corollary 12.1.1, the following equality holds

$$\max_{\phi^{AA'^n}} I(A;B^n)_\omega = n\ I(\mathcal{N}). \tag{20.127}$$

Thus, the bound on the classical rate C of a reliable protocol for entanglement-assisted common randomness generation is

$$nC \leq n\ I(\mathcal{N}) + n\epsilon', \tag{20.128}$$

and it also serves as an upper bound for entanglement-assisted classical communication. This demonstrates a single-letter upper bound on the entanglement-assisted classical capacity of a quantum channel and completes the proof of Theorem 20.3.1.

20.5.1 Feedback Does Not Increase Capacity

The entanglement-assisted classical capacity formula is the closest formal analogy to Shannon's capacity formula for a classical channel. The mutual information $I(\mathcal{N})$ of a quantum channel $\mathcal{N}$ is the optimum of the quantum mutual information over all bipartite input states:

$$I(\mathcal{N}) = \max_{\phi^{AA'}} I(A;B), \tag{20.129}$$

and it is equal to the channel's entanglement-assisted classical capacity by Theorem 20.4.1. The mutual information $I(p_{Y|X})$ of a classical channel $p_{Y|X}$ is the optimum of the classical mutual information over all correlated inputs to the channel:

$$I(p_{Y|X}) = \max_{XX'} I(X;Y), \tag{20.130}$$

where XX' are correlated random variables with the distribution $p_{X,X'}(x,x') = p_X(x)\delta_{x,x'}$. The formula is equal to the classical capacity of a classical channel by Shannon's noisy coding theorem. Both formulas not only appear similar in form, but they also have the important property of being "single-letter," meaning that the above formulas are equal to the capacity (this was not the case for the Holevo information from the previous chapter).

We now consider another way in which the entanglement-assisted classical capacity is the best candidate for being the generalization of Shannon's formula to the quantum world. Though it might be surprising, it is well known that free access to a classical feedback channel from receiver to sender does not increase the capacity of a classical channel. We state this result as the following theorem.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Theorem 20.5.1 (Feedback does not increase classical capacity)**.** *The feedback capacity of a classical channel $p_{Y|X}(y|x)$ is equal to the mutual information of that channel:*

$$\sup\{C : C \text{ is achievable with feedback } \} = I\big(p_{Y|X}\big), \tag{20.131}$$

where $I\big(p_{Y|X}\big)$ is defined in (20.130).

Proof. We first define an $(n, C-\delta, \epsilon)$ classical feedback code as one in which every symbol $x_i(m, Y^{i-1})$ of a codeword $x^n(m)$ is a function of the message $m \in \mathcal{M}$ and all of the previous received values $Y_1, \ldots, Y_{i-1}$ from the receiver. The decoder consists of the decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, |\mathcal{M}|\}$ such that

$$\Pr\{M' \neq M\} \leq \epsilon, \tag{20.132}$$

where $M' \equiv g(Y^n)$. The lower bound LHS $\geq$ RHS follows because we can always avoid the use of the feedback channel and achieve the mutual information of the classical channel by employing Shannon's noisy coding theorem. The upper bound LHS $\leq$ RHS is less obvious, but it follows from the memoryless structure of the channel and the structure of a feedback code. Consider the following chain of inequalities:

$$\begin{aligned}
nC &= H(M) && (20.133)\\
&= I(M; M') + H(M \mid M') && (20.134)\\
&\leq I(M; M') + 1 + \epsilon nC && (20.135)\\
&\leq I(M; Y^n) + 1 + \epsilon nC. && (20.136)
\end{aligned}$$

The first equality follows because we assume that the message M is uniformly distributed. The first inequality follows from Fano's inequality (see Theorem 10.7.3) and the assumption in (20.132) that the protocol is good up to error ϵ. The last inequality follows from classical data processing. Continuing, we can bound $I(M; Y^n)$ from above:

$$\begin{aligned}
I(M; Y^n) &= H(Y^n) - H(Y^n|M) && (20.137)\\
&= H(Y^n) - \sum_{k=1}^{n} H\big(Y_k \mid Y^{k-1}M\big) && (20.138)\\
&= H(Y^n) - \sum_{k=1}^{n} H\big(Y_k \mid Y^{k-1}MX_k\big) && (20.139)\\
&= H(Y^n) - \sum_{k=1}^{n} H(Y_k \mid X_k) && (20.140)\\
&\leq \sum_{k=1}^{n} H(Y_k) - H(Y_k \mid X_k) && (20.141)\\
&= \sum_{k=1}^{n} I(X_k; Y_k) && (20.142)\\
&\leq n \max_{XX'} I(X; Y) && (20.143)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first equality follows from the definition of mutual information. The second equality follows from the chain rule for entropy (see Exercise 10.3.2). The third equality follows because X_k is a function of Y^{k-1} and M. The fourth equality follows because Y_k is conditionally independent of Y^{k-1} and M through X_k ($Y^{k-1}M \to X_k \to Y_k$ forms a Markov chain). The first inequality follows from subadditivity of entropy. The fifth equality follows by definition, and the final inequality follows because the individual mutual informations in the sum can never exceed the maximum over all inputs. Putting everything together, our final bound on the feedback-assisted capacity of a classical channel is

$$C \leq I\left(p_{Y|X}\right) + \frac{1}{n} + \epsilon C, \tag{20.144}$$

which becomes $C \leq I\left(p_{Y|X}\right)$ as $n \to \infty$ and $\epsilon \to 0$. □

Given the above result, we might wonder if a similar result could hold for the entanglement-assisted classical capacity. Such a result would more firmly place the entanglement-assisted classical capacity as a good generalization of Shannon's coding theorem. Indeed, the following theorem states that this result holds.

Theorem 20.5.2 (Quantum feedback does not increase the EAC capacity)**.** *The classical capacity of a quantum channel assisted by a quantum feedback channel is equal to that channel's entanglement-assisted classical capacity:*

$$\sup\{C \mid C \textit{ is achievable with quantum feedback}\} = I(\mathcal{N}), \tag{20.145}$$

where $I(\mathcal{N})$ is defined in (20.129).

Proof. We define free access to a quantum feedback channel to mean that there is a noiseless quantum channel of arbitrarily large dimension going from the receiver Bob to the sender Alice. The bound LHS $\geq$ RHS follows because Bob can use the quantum feedback channel to establish an arbitrarily large amount of entanglement with Alice. They then just execute the protocol from Section 20.4 to achieve a rate equal to the entanglement-assisted classical capacity. The bound LHS $\leq$ RHS is much less obvious, and it requires a proof that is different from the proof of Theorem 20.5.1. We first need to determine the most general protocol for classical communication with the assistance of a quantum feedback channel. Figure 20.4 depicts such a protocol with Alice's systems in red, Bob's systems in blue, and the feedback systems in green. Alice begins by correlating the message M with quantum codewords of the form $\rho_m^{A^n}$, where the state $\rho_m^{A^n}$ can be entangled across all of the channel uses. We describe the k^{th} step of the protocol with quantum feedback:

1. Bob receives the channel output B_k. He acts with some unitary U^k on B_k, his previously received systems B^{k-1}, and two new registers X_k and Y_k. We place no restriction on the size of the registers X_k and Y^k.

2. Bob transmits the system X_k through the noiseless feedback channel to Alice.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

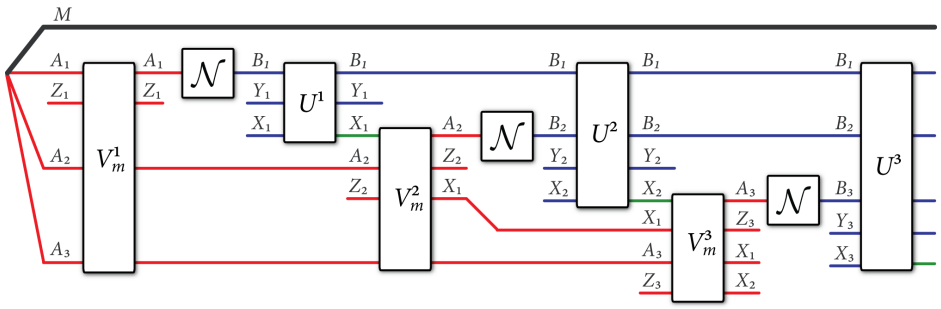

Figure 20.4: Three rounds of the most general protocol for classical communication with a quantum feedback channel. Alice's systems are in red, Bob's are in blue, and quantum feedback systems are in green. Alice has a message M classically correlated with the quantum systems $A_1A_2A_3$. She acts with a unitary V_m^1 on $A_1A_2A_3$ and an ancilla Z_1, depending on her message m. She transmits A_1 through the noisy channel. Bob receives B_1 from the channel and performs a unitary U^1 on B_1, and two other registers X_1 and Y_1. He sends X_1 through the quantum feedback channel to Alice, and Alice continues the encoding by exploiting the feedback system. This process continues in the k^{th} round with Bob performing unitaries on his previous systems B^k, and Alice performing unitaries on her current register $A_k \cdots A_n$ and her feedback registers X^{k-1}.

3. Alice acts with a unitary V_m^k that depends on the message m. This unitary acts on her received system X_k, all of the other systems $A_k \cdots A_n$ in her possession, an ancilla Z_k, and all of her previously processed feedback systems X^{k-1}.

4. She transmits the system A_k through the noisy quantum channel.

This protocol is the most general for classical communication with quantum feedback because all of its operations are inclusive of previous steps in the protocol. Also, the most general operations could have CPTP maps rather than unitaries, but Alice and Bob can discard their ancilla systems in order to simulate such maps. We can now proceed with proving the upper bound LHS $\leq$ RHS. To do so, we assume that the random variable M modeling Alice's message selection is a uniform random variable, and Bob obtains a random variable M' by measuring all of his systems B^n at the end of the protocol. For any good protocol for classical communication, the bound $\Pr\{M' \neq M\} \leq \epsilon$ applies. Consider the following chain of inequalities (these steps are essentially the same as those in (20.133-20.136)):

$$\begin{aligned} nC &= H(M) && (20.146)\\ &= I(M;M') + H(M \mid M') && (20.147)\\ &\leq I(M;M') + 1 + \epsilon nC && (20.148)\\ &\leq I(M;B^n) + 1 + \epsilon nC && (20.149) \end{aligned}$$

This chain of inequalities follows for the same reason as those in (20.133-20.136), with the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

last step following from quantum data processing. Continuing, we have

$$I(M;B^n) = I\big(M;B_n|B^{n-1}\big) + I\big(M;B^{n-1}\big) \tag{20.150}$$
$$\leq I\big(M;B_n|B^{n-1}\big) + I\big(M;B_{n-1}|B^{n-2}\big) + I\big(M;B^{n-2}\big) \tag{20.151}$$
$$\leq \sum_{k=1}^{n} I\big(M;B_k|B^{k-1}\big) \tag{20.152}$$
$$\leq n \max_{\rho} I(M;B|A) \tag{20.153}$$

The first equality follows from the chain rule for quantum mutual information. The first inequality follows from quantum data processing and another application of the chain rule. The third inequality follows from recursively applying the same inequality. The final inequality follows from considering that the state on systems M, B_k, and B^{k-1} is a particular state of the form:

$$\rho^{MBA} \equiv \sum_m p_M(m)|m\rangle\langle m|^M \otimes \mathcal{N}^{A'\to B}(\rho_m^{A'A}), \tag{20.154}$$

and so $I(M;B_k|B^{n-1})$ can never be greater than the maximization over all states of this form. We can then bound $\max_\rho I(M;B|A)$ by the quantum mutual information $I(\mathcal{N})$ of the channel:

$$I(M;B|A) = H(B|A) - H(B|AM) \tag{20.155}$$
$$\leq H(B) - \sum_m p_M(m) H(B|A)_{\rho_m} \tag{20.156}$$
$$= H(B) + \sum_m p_M(m) H(B|E)_{\psi_m} \tag{20.157}$$
$$= H(B) + H(B|EM)_{\psi^{MBE}} \tag{20.158}$$
$$\leq H(B) + H(B|E)_{\psi} \tag{20.159}$$
$$= H(B) - H(B|A)_{\phi} \tag{20.160}$$
$$\leq I(\mathcal{N}) \tag{20.161}$$

The first equality follows from expanding the conditional quantum mutual information. The first inequality follows from subadditivity of entropy and expanding the conditional entropy $H(B|AM)$ with the classical variable M. The second equality follows by taking ψ_m^{BAE} as a purification of the state ρ_m^{BA} and considering that $-H(B|A) = H(B|E)$ for any tripartite pure state. The third equality follows by rewriting the convex sum of entropies as a conditional entropy with the classical system M and the state $\psi^{MBAE} = \sum_m p(m)|m\rangle\langle m| \otimes \psi_m^{BAE}$. The second inequality follows because conditioning cannot increase entropy, and the fourth equality follows by taking ϕ^{ABE} as a purification of ψ^{BE}. The final inequality follows by noting that $H(B) - H(B|A) = I(A;B)$ and this quantum mutual information can never be greater than the maximum. Putting everything together, we get the following upper bound on any achievable rate C for classical communication with quantum feedback:

$$C \leq I(\mathcal{N}) + \frac{1}{n} + \epsilon C, \tag{20.162}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

which becomes $C \leq I(\mathcal{N})$ as $n \to \infty$ and $\epsilon \to 0$. □

Corollary 20.5.1. *The capacity of a quantum channel with unlimited entanglement and classical feedback is equal to the entanglement-assisted classical capacity of $\mathcal{N}$.*

Proof. This result follows because we have that $I(\mathcal{N})$ is a lower bound on this capacity (simply by avoiding use of the classical feedback channel). Also, $I(\mathcal{N})$ is an upper bound on this capacity because the entanglement and classical feedback channel can simulate an arbitrarily large quantum feedback channel via teleportation, and the above theorem gives an upper bound of $I(\mathcal{N})$ for this setting. □

20.6 Examples of Channels

This section shows how to compute the entanglement-assisted classical capacity of both the quantum erasure channel and the amplitude damping channel, while leaving the capacity of the quantum depolarizing channel and the dephasing channel as exercises. For three of these channels (erasure, depolarizing, and dephasing), a super-dense-coding-like strategy suffices to achieve capacity. This strategy involves Alice locally rotating an ebit shared with Bob, sending half of it through the noisy channel, and Bob performing measurements in the Bell basis to determine what Alice sent. This process induces a classical channel from Alice to Bob, for which its capacity is equal to the entanglement-assisted capacity of the original quantum channel (in the case of depolarizing, dephasing, and erasure channels). For the amplitude damping channel, this super-dense-coding-like strategy does not achieve capacity—in general, it is necessary for Bob to perform a large, collective measurement on all of the channel outputs in order for him to determine Alice's message.

Figure 20.5 plots the entanglement-assisted capacities of these four channels as a function of their noise parameters. As expected, the depolarizing channel has the worst performance because it is a "worst-case scenario" channel—it either sends the state through or replaces it with a completely random state. The erasure channel's capacity is just a line of constant slope down to zero—this is because the receiver can easily determine the fraction of the time that he receives something from the channel. The dephasing channel eventually becomes a completely classical channel, for which entanglement cannot increase capacity beyond one bit per channel use. Finally, perhaps the most interesting curve is for the amplitude damping channel. This channel's capacity is convex when its noise parameter is less than 1/2 and concave when it is greater than 1/2.

20.6.1 The Quantum Erasure Channel

Recall that the quantum erasure channel acts as follows on an input density operator $\rho^{A'}$:

$$\rho^{A'} \to (1-\epsilon)\rho^{B} + \epsilon |e\rangle\langle e|^{B}, \tag{20.163}$$

where ϵ is the erasure probability and $|e\rangle^{B}$ is an erasure state that is orthogonal to the support of the input state ρ.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

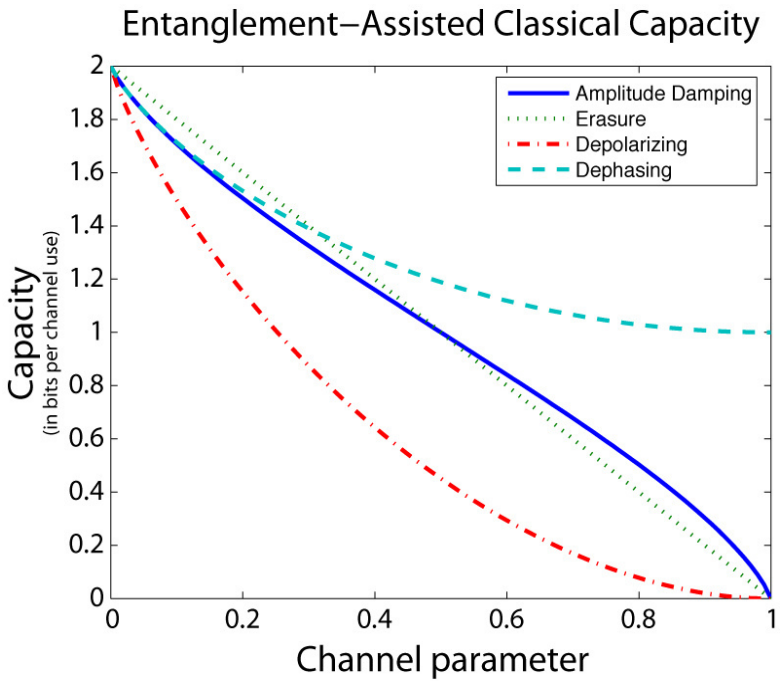

Figure 20.5: The entanglement-assisted classical capacity of the amplitude damping channel, the erasure channel, the depolarizing channel, and the dephasing channel as a function of each channel's noise parameter.

Proposition 20.6.1. *The entanglement-assisted classical capacity of a quantum erasure channel with erasure probability ϵ is*

$$2(1-\epsilon)\log d_A, \tag{20.164}$$

where d_A is the dimension of the input system.

Proof. To determine the entanglement-assisted classical capacity of this channel, we need to compute its mutual information. So, consider that sending half of a bipartite state $\phi^{AA'}$ through the channel produces the output

$$\sigma^{AB} \equiv (1-\epsilon)\phi^{AB} + \epsilon\phi^{A} \otimes |e\rangle\langle e|^{B}. \tag{20.165}$$

We could now attempt to calculate and optimize the quantum mutual information $I(A;B)_{\sigma}$. Though, observe that Bob can apply the following isometry $U^{B\to BX}$ to his state:

$$U^{B\to BX} \equiv \Pi^{B} \otimes |0\rangle^{X} + |e\rangle\langle e|^{B} \otimes |1\rangle^{X}, \tag{20.166}$$

where Π^{B} is a projector onto the support of the input state (for qubits, it would be just $|0\rangle\langle 0| + |1\rangle\langle 1|$). Applying this isometry leads to a state σ^{ABX} where

$$\begin{aligned} \sigma^{ABX} &\equiv U^{B\to BX}\sigma^{AB}\left(U^{B\to BX}\right)^{\dagger} && (20.167)\\ &= (1-\epsilon)\phi^{AB} \otimes |0\rangle\langle 0|^{X} + \epsilon\phi^{A} \otimes |e\rangle\langle e|^{B} \otimes |1\rangle\langle 1|^{X}. && (20.168)\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The quantum mutual information $I(A;BX)_\sigma$ is equal to $I(A;B)_\sigma$ because entropies do not change under the isometry $U^{B\to BX}$. We now calculate $I(A;BX)_\sigma$:

$$\begin{aligned}
I(A;BX)_\sigma &= H(A)_\sigma + H(BX)_\sigma - H(ABX)_\sigma && (20.169)\\
&= H(A)_\phi + H(B|X)_\sigma - H(AB|X)_\sigma && (20.170)\\
&= H(A)_\phi + (1-\epsilon)\Big[H(B)_\phi - H(AB)_\phi\Big] \\
&\quad + \epsilon\Big[H(B)_{|e\rangle} - H(AB)_{\phi^A\otimes|e\rangle\langle e|}\Big] && (20.171)\\
&= H(A)_\phi + (1-\epsilon)H(B)_\phi - \epsilon\Big[H(A)_\phi + H(B)_{|e\rangle}\Big] && (20.172)\\
&= (1-\epsilon)\Big[H(A)_\phi + H(B)_\phi\Big] && (20.173)\\
&= 2(1-\epsilon)H(A)_\phi && (20.174)\\
&\leq 2(1-\epsilon)\log d_A. && (20.175)
\end{aligned}$$

The first equality follows by the definition of quantum mutual information. The second equality follows from $\phi^A = \mathrm{Tr}_{BX}\{\sigma^{ABX}\}$, from the chain rule of entropy, and by canceling $H(X)$ on both sides. The third equality follows because the X register is a classical register, indicating whether the erasure occurs. The fourth equality follows because $H(AB)_\phi = 0$, $H(B)_{|e\rangle} = 0$, and $H(AB)_{\phi^A\otimes|e\rangle\langle e|} = H(A)_\phi + H(B)_{|e\rangle}$. The fifth equality follows again because $H(B)_{|e\rangle} = 0$ and by collecting terms. The final equality follows because $H(A)_\phi = H(B)_\phi$ (ϕ^{AB} is a pure bipartite state). The final inequality follows because the entropy of a state on system A is never greater than logarithm of the dimension of A. We can conclude that the maximally entangled state $\Phi^{AA'}$ achieves the entanglement-assisted classical capacity of the quantum erasure channel because $H(A)_\Phi = \log d_A$. □

The strategy for achieving the entanglement-assisted classical capacity of the quantum erasure channel is straightforward. Alice and Bob simply employ a super-dense coding strategy on all of the channel uses (this means that Bob performs measurements on each channel output with his share of the entanglement—there is no need for a large, collective measurement on all of the channel outputs). For a good fraction $1-\epsilon$ of the time, this strategy works and Alice can communicate $2\log d_A$ bits to Bob. For the other fraction ϵ, all is lost to the environment. In order for this to work, Alice and Bob need to make use of a feedback channel from Bob to Alice so that Bob can report which messages come through and which do not, but Corollary 20.5.1 states that this feedback cannot improve the capacity. Thus, the rate of communication they can achieve is equal to the capacity $2(1-\epsilon)\log d_A$.

20.6.2 The Amplitude Damping Channel

We now compute the entanglement-assisted classical capacity of the amplitude damping channel $\mathcal{N}_{\mathrm{AD}}$. Recall that this channel acts as follows on an input qubit in state ρ:

$$\mathcal{N}_{\mathrm{AD}}(\rho) = A_0\rho A_0^\dagger + A_1\rho A_1^\dagger, \tag{20.176}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where

$$A_0 \equiv |0\rangle\langle 0| + \sqrt{1-\gamma}|1\rangle\langle 1|, \qquad A_1 \equiv \sqrt{\gamma}|0\rangle\langle 1|. \tag{20.177}$$

Proposition 20.6.2. *The entanglement-assisted classical capacity of an amplitude damping channel with damping parameter γ is*

$$I(\mathcal{N}_{AD}) = \max_{p\in[0,1]} H_2(p) + H_2((1-\gamma)p) - H_2(\gamma p), \tag{20.178}$$

where $H_2(p) \equiv -p\log p - (1-p)\log(1-p)$ is the binary entropy function.

Proof. Suppose that a matrix representation of the input qubit density operator ρ in the computational basis is

$$\rho = \begin{bmatrix} 1-p & \eta^* \\ \eta & p \end{bmatrix}. \tag{20.179}$$

One can readily verify that the density operator for Bob has the following matrix representation:

$$\mathcal{N}_{\text{AD}}(\rho) = \begin{bmatrix} 1-(1-\gamma)p & \sqrt{1-\gamma}\eta^* \\ \sqrt{1-\gamma}\eta & (1-\gamma)p \end{bmatrix}, \tag{20.180}$$

and by calculating the elements $\text{Tr}\{A_i \rho A_j^\dagger\}|i\rangle\langle j|$, we can obtain a matrix representation for Eve's density operator:

$$\mathcal{N}^c_{\text{AD}}(\rho) = \begin{bmatrix} 1-\gamma p & \sqrt{\gamma}\eta^* \\ \sqrt{\gamma}\eta & \gamma p \end{bmatrix}, \tag{20.181}$$

where $\mathcal{N}^c_{\text{AD}}$ is the complementary channel to Eve. By comparing (20.180) and (20.181), we can see that the channel to Eve is an amplitude damping channel with damping parameter $1-\gamma$. The entanglement-assisted classical capacity of $\mathcal{N}_{\text{AD}}$ is equal to its mutual information:

$$I(\mathcal{N}_{\text{AD}}) = \max_{\phi^{AA'}} I(A;B)_\sigma, \tag{20.182}$$

where $\phi^{AA'}$ is some pure bipartite input state and $\sigma^{AB} = \mathcal{N}_{\text{AD}}(\phi^{AA'})$. We need to determine the input density operator that maximizes the above formula as a function of γ. As it stands now, the optimization depends on three parameters: p, $\text{Re}\{\eta\}$, and $\text{Im}\{\eta\}$. We can show that it is sufficient to consider an optimization over only p with $\eta = 0$. The formula in (20.182) also has the following form:

$$I(\mathcal{N}_{\text{AD}}) = \max_{\rho} [H(\rho) + H(\mathcal{N}_{\text{AD}}(\rho)) - H(\mathcal{N}^c_{\text{AD}}(\rho))], \tag{20.183}$$

because

$$\begin{aligned} I(A;B)_\sigma &= H(A)_\phi + H(B)_\sigma - H(AB)_\sigma && (20.184)\\ &= H(A')_\phi + H(\mathcal{N}_{\text{AD}}(\rho)) - H(E)_\sigma && (20.185)\\ &= H(\rho) + H(\mathcal{N}_{\text{AD}}(\rho)) - H(\mathcal{N}^c_{\text{AD}}(\rho)) && (20.186)\\ &\equiv I_{\text{mut}}(\rho, \mathcal{N}_{\text{AD}}). && (20.187) \end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The three entropies in (20.183) depend only on the eigenvalues of the three density operators in (20.179-20.181), respectively, which are as follows:

$$\frac{1}{2}\left(1 \pm \sqrt{(1-2p)^2 + 4|\eta|^2}\right), \tag{20.188}$$

$$\frac{1}{2}\left(1 \pm \sqrt{(1-2(1-\gamma)p)^2 + 4|\eta|^2(1-\gamma)}\right), \tag{20.189}$$

$$\frac{1}{2}\left(1 \pm \sqrt{(1-2\gamma p)^2 + 4|\eta|^2\gamma}\right). \tag{20.190}$$

The above eigenvalues are in the order of Alice, Bob, and Eve. All of the above eigenvalues have a similar form, and their dependence on η is only through its magnitude. Thus, it suffices to consider $\eta \in \mathbb{R}$ (this eliminates one parameter). Next, the eigenvalues do not change if we flip the sign of η (this is equivalent to rotating the original state ρ by Z, to $Z\rho Z$), and thus, the mutual information does not change as well:

$$I_{\text{mut}}(\rho, \mathcal{N}_{\text{AD}}) = I_{\text{mut}}(Z\rho Z, \mathcal{N}_{\text{AD}}). \tag{20.191}$$

By the above relation and concavity of quantum mutual information in the input density operator (Theorem 12.4.2), the following inequality holds

$$I_{\text{mut}}(\rho, \mathcal{N}_{\text{AD}}) = \frac{1}{2}[I_{\text{mut}}(\rho, \mathcal{N}_{\text{AD}}) + I_{\text{mut}}(Z\rho Z, \mathcal{N}_{\text{AD}})] \tag{20.192}$$

$$\leq I_{\text{mut}}(\frac{1}{2}(\rho + Z\rho Z), \mathcal{N}_{\text{AD}}) \tag{20.193}$$

$$= I_{\text{mut}}(\overline{\Delta}(\rho), \mathcal{N}_{\text{AD}}), \tag{20.194}$$

where $\overline{\Delta}$ is a completely dephasing channel in the computational basis. This demonstrates that it is sufficient to consider diagonal density operators ρ when optimizing the quantum mutual information. Thus, the eigenvalues in (20.188-20.190) respectively become

$$\{p, 1-p\}, \tag{20.195}$$

$$\{(1-\gamma)p, 1-(1-\gamma)p\}, \tag{20.196}$$

$$\{\gamma p, 1-\gamma p\}, \tag{20.197}$$

giving our final expression in the statement of the proposition. □

Exercise 20.6.1 Consider the qubit depolarizing channel: $\rho \to (1-p)\rho + p\pi$. Prove that its entanglement-assisted classical capacity is equal to

$$2 + (1 - 3p/4)\log(1 - 3p/4) + (3p/4)\log(p/4). \tag{20.198}$$

Exercise 20.6.2 Consider the dephasing channel: $\rho \to (1-p/2)\rho + (p/2)Z\rho Z$. Prove that its entanglement-assisted classical capacity is equal to $2 - H_2(p/2)$, where p is the dephasing parameter.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

20.7 Concluding Remarks

Shared entanglement has the desirable property of simplifying quantum Shannon theory. The entanglement-assisted capacity theorem is one of the strongest known results in quantum Shannon theory because it states that the quantum mutual information of a channel is equal to its entanglement-assisted capacity. This function of the channel is concave in the input state and the set of input states is convex, implying that finding a local maximum is equivalent to finding a global one. The converse theorem demonstrates that there is no need to take the regularization of the formula—strong subadditivity guarantees that it is additive. Furthermore, feedback does not improve this capacity, just as it does not for the classical case of Shannon's setting. In these senses, the entanglement-assisted classical capacity is the most natural generalization of Shannon's capacity formula to the quantum setting.

The direct coding part of the capacity theorem exploits a strategy similar to super-dense coding—effectively the technique is to perform super-dense coding in the type class subspaces of many copies of a shared entangled state. This strategy is *equivalent* to super-dense coding if the initial shared state is a maximally entangled state. The particular protocol that we outlined in this chapter has the appealing feature that we can easily make it coherent, similar to the way that coherent dense coding is a coherent version of the super-dense coding protocol. We take this approach in the next chapter and show that we can produce a whole host of other protocols with this technique, eventually leading to a proof of the direct coding part of the quantum capacity theorem.

This chapter features the calculation of the entanglement-assisted classical capacity of certain channels of practical interest: the depolarizing channel, the dephasing channel, the amplitude damping channel, and the erasure channel. Each one of these channels has a single parameter that governs its noisiness, and the capacity in each case is a straightforward function of this parameter. One could carry out a similar type of analysis to determine the entanglement-assisted capacity of any channel, although it generally will be necessary to employ techniques from convex optimization.

Unfortunately, quantum Shannon theory only gets more complicated from here onward.[3] For the other capacity theorems that we will study, such as the private classical capacity or the quantum capacity, the best expressions that we have for them are good only up to regularization of the formulas. In certain cases, these formulas completely characterize the capabilities of the channel for these particular operational tasks, but these formulas are not particularly useful in the general case. One important goal for future research in quantum Shannon theory would be to improve upon these formulas, in the hopes that we could further our understanding of the best strategy for achieving the information processing tasks corresponding to these other capacity questions.

[3] We could also view this "unfortunate" situation as being fortunate for conducting open-ended research in quantum Shannon theory.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

20.8 History and Further Reading

Adami and Cerf figured that the mutual information of a quantum channel would play an important role in quantum Shannon theory, and they proved several of its most important properties [5]. Bennett *et al.* later demonstrated that the quantum mutual information of a channel has the operational interpretation as its entanglement-assisted classical capacity [33, 34]. Our proof of the direct part of the entanglement-assisted classical capacity theorem is the same as that in Ref. [156]. We exploit this approach because it leads to all of the results in the next chapter, implying that this protocol is sufficient to generate all of the known protocols in quantum Shannon theory (with the exception of private classical communication). Giovanetti and Fazio determined several capacities of the amplitude damping channel [102], and Perez-Garcia and Wolf made some further observations regarding it [262]. Bowen *et al.* proved that the classical capacity of a channel assisted by unbounded quantum feedback is equal to its entanglement-assisted classical capacity [42, 44, 43].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 21

Coherent Communication with Noisy Resources

This chapter demonstrates the power of both coherent communication from Chapter 7 and the particular protocol for entanglement-assisted classical coding from the previous chapter. Recall that coherent dense coding is a version of the dense coding protocol in which the sender and receiver perform all of its steps coherently.[1] Since our protocol for entanglement-assisted classical coding from the previous chapter is really just a glorified dense coding protocol, the sender and receiver can perform each of its steps coherently, generating a protocol for entanglement-assisted coherent coding. Then, by exploiting the fact that two coherent bits are equivalent to a qubit and an ebit, we obtain a protocol for entanglement-assisted quantum coding that consumes far less entanglement than a naive strategy would in order to accomplish this task. We next combine this entanglement-assisted quantum coding protocol with entanglement distribution (Section 6.2.1) and obtain a protocol for which the channel's coherent information (Section 12.5) is an achievable rate for quantum communication. This sequence of steps demonstrates an alternate proof of the direct part of the quantum channel coding theorem in Chapter 23.

Entanglement-assisted classical communication is one generalization of super-dense coding, in which the noiseless qubit channel becomes an arbitrary noisy quantum channel while the noiseless ebits remain noiseless. Another generalization of super-dense coding is a protocol named *noisy super-dense coding*, in which the shared entanglement becomes a shared noisy state ρ^{AB} and the noiseless qubit channels remain noiseless. Interestingly, the protocol that we employ in this chapter for noisy super-dense coding is essentially equivalent to the protocol from the previous chapter for entanglement-assisted classical communication, with some slight modifications to account for the different setting. We can also construct a coherent version of noisy super-dense coding, leading to a protocol that we name *coherent state transfer*. Coherent state transfer accomplishes not only the task of generating coherent communication between Alice and Bob, but it also allows Alice to transfer her share of the

[1] Performing a protocol coherently means that we replace conditional unitaries with controlled unitaries and measurements with controlled gates (e.g., see Figures 6.2 and 7.3).

state ρ^{AB} to Bob. By combining coherent state transfer with both the coherent communication identity and teleportation, we obtain protocols for quantum-assisted state transfer and classical-assisted state transfer, respectively. The latter protocol gives an operational interpretation to the quantum conditional entropy $H(A|B)_\rho$—if it is positive, then the protocol consumes entanglement at the rate $H(A|B)_\rho$, and if it is negative, the protocol generates entanglement at the rate $|H(A|B)_\rho|$.

The final part of this chapter shows that our particular protocol for entanglement-assisted classical communication is even more powerful than suggested in the first paragraph. It allows for a sender to communicate both coherent bits and incoherent classical bits to a receiver, and they can trade off these two resources against one another. The structure of the entanglement-assisted protocol allows for this possibility, by taking advantage of Remark 20.4.1 and by combining it with the HSW classical communication protocol from Chapter 19. Then, by exploiting the coherent communication identity, we obtain a protocol for entanglement-assisted communication of classical and quantum information. Chapter 24 demonstrates that this protocol, teleportation, super-dense coding, and entanglement distribution are sufficient to accomplish any task in dynamic quantum Shannon theory involving the three unit resources of classical bits, qubits, and ebits. These four protocols give a three-dimensional achievable rate region that is the best known characterization for any information processing task that a sender and receiver would like to accomplish with a noisy channel and the three unit resources. Chapter 24 discusses this triple trade-off scenario in full detail.

21.1 Entanglement-Assisted Quantum Communication

The entanglement-assisted classical capacity theorem states that the quantum mutual information of a channel is equal to its capacity for transmitting classical information with the help of shared entanglement, and the direct coding theorem from Section 20.4 provides a protocol that achieves the capacity. We were not much concerned with the rate at which this protocol consumes entanglement, but a direct calculation reveals that it consumes $H(A)_\varphi$ ebits per channel use, where $|\varphi\rangle^{AB}$ is the bipartite state that they share before the protocol begins.[2]

Suppose now that Alice is interested in exploiting the channel and shared entanglement in order to transmit quantum information to Bob. There is a simple (and as we will see, naive) way that we can convert the protocol in Section 20.4 to one that transmits quantum information: they can just combine it with teleportation. This naive strategy requires consuming ebits at an additional rate of $\frac{1}{2}I(A;B)_\rho$ in order to have enough entanglement to combine with teleportation, where $\rho^{AB} \equiv \mathcal{N}^{A'\to B}(\varphi^{AA'})$. To see this, consider the following

[2]This result follows because they can concentrate n copies of the state $|\varphi\rangle^{AB}$ to $nH(A)_\varphi$ ebits, as we learned in Chapter 18. Also, they can "dilute" $nH(A)_\varphi$ ebits to n copies of $|\varphi\rangle^{AB}$ with the help of a sublinear amount of classical communication that does not factor into the resource count (we have not studied the protocol for entanglement dilution).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

resource inequalities:

$$\langle \mathcal{N} \rangle + \left(H(A)_\rho + \frac{1}{2} I(A;B)_\rho \right) [qq] \geq I(A;B)_\rho [c \rightarrow c] + \frac{1}{2} I(A;B)_\rho [qq] \tag{21.1}$$

$$\geq \frac{1}{2} I(A;B)_\rho [q \rightarrow q]. \tag{21.2}$$

The first inequality follows by having them exploit the channel and the $nH(A)_\rho$ ebits to generate classical communication at a rate $I(A;B)_\rho$ (while doing nothing with the extra $n\frac{1}{2}I(A;B)_\rho$ ebits). Alice then exploits the ebits and the classical communication in a teleportation protocol to send $n\frac{1}{2}I(A;B)_\rho$ qubits to Bob. This rate of quantum communication is provably optimal—were it not so, it would be possible to combine the protocol in (21.1-21.2) with super-dense coding and beat the optimal rate for classical communication given by the entanglement-assisted classical capacity theorem.

Although the above protocol achieves the entanglement-assisted quantum capacity, we are left thinking that the entanglement consumption rate of $H(A)_\rho + \frac{1}{2}I(A;B)_\rho$ ebits per channel use might be a bit more than necessary because teleportation and super-dense coding are not dual under resource reversal. That is, if we combine the protocol with super-dense coding and teleportation *ad infinitum*, then it consumes an infinite amount of entanglement. In practice, this "back and forth" with teleportation and super-dense coding would be a poor way to consume the precious resource of entanglement.

How might we make more judicious use of shared entanglement? Recall that coherent communication from Chapter 7 was helpful for doing so, at least in the noiseless case. A sender and receiver can combine coherent teleportation and coherent dense coding *ad infinitum* without any net loss in entanglement, essentially because these two protocols are dual under resource reversal. The following theorem shows how we can upgrade the protocol in Section 20.4 to one that generates coherent communication instead of just classical communication. The resulting protocol is one way to have a version of coherent dense coding in which one noiseless resource is replaced by a noisy one.

Theorem 21.1.1 (Entanglement-Assisted Coherent Communication)**.** *The following resource inequality corresponds to an achievable protocol for entanglement-assisted coherent communication over a noisy quantum channel:*

$$\langle \mathcal{N} \rangle + H(A)_\rho [qq] \geq I(A;B)_\rho [q \rightarrow qq], \tag{21.3}$$

where $\rho^{AB} \equiv \mathcal{N}^{A' \rightarrow B}(\varphi^{AA'})$.

Proof. Suppose that Alice and Bob share many copies of some pure, bipartite entangled state $|\varphi\rangle^{AB}$. Consider the code from the direct coding theorem in Section 20.4. We can say that it is a set of $D^2 \approx 2^{nI(A;B)_\rho}$ unitaries $U(s(m))$, from which Alice can select, and she applies a particular unitary $U(s(m))$ to her share A^n of the entanglement in order to encode message m. Also, Bob has a detection POVM $\left\{ \Lambda_m^{B'^n B^n} \right\}$ acting on his share of the entanglement and the channel outputs that he can exploit to detect message m. Just as we

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

were able to construct a coherent super-dense coding protocol in Chapter 7 by performing all the steps in dense coding coherently, we can do so for the entanglement-assisted classical coding protocol in Section 20.4. We track the steps in such a protocol. Suppose Alice shares a state with a reference system R to which she does not have access:

$$|\psi\rangle^{RA_1} \equiv \sum_{l,m=1}^{D^2} \alpha_{l,m} |l\rangle^R |m\rangle^{A_1}, \tag{21.4}$$

where $\{|l\rangle\}$ and $\{|m\rangle\}$ are some orthonormal bases for R and A_1, respectively. We say that Alice and Bob have implemented a coherent channel if they execute the map $|m\rangle^{A_1} \rightarrow |m\rangle^{A_1}|m\rangle^{B_1}$, which transforms the above state to

$$\sum_{l,m=1}^{D^2} \alpha_{l,m} |l\rangle^R |m\rangle^{A_1} |m\rangle^{B_1}. \tag{21.5}$$

We say that they have implemented a coherent channel *approximately* if the state resulting from the protocol is ϵ-close in trace distance to the above state. If we can show that ϵ is an arbitrary positive number that approaches zero in the asymptotic limit, then the simulation of an approximate coherent channel asymptotically becomes an exact simulation. Alice's first step is to append her shares of the entangled state $|\varphi\rangle^{A^n B^n}$ to $|\psi\rangle^{RA_1}$ and apply the following controlled unitary from her system A_1 to her system A^n:

$$\sum_m |m\rangle\langle m|^{A_1} \otimes U(s(m))^{A^n}. \tag{21.6}$$

The resulting global state is as follows:

$$\sum_{l,m} \alpha_{l,m} |l\rangle^R |m\rangle^{A_1} U(s(m))^{A^n} |\varphi\rangle^{A^n B^n}. \tag{21.7}$$

By the structure of the unitaries $U(s(m))$ (see (20.46) and (20.48)), the above state is equivalent to the following one:

$$\sum_{l,m} \alpha_{l,m} |l\rangle^R |m\rangle^{A_1} \left(U^T(s(m))\right)^{B^n} |\varphi\rangle^{A^n B^n}. \tag{21.8}$$

Interestingly, observe that Alice applying the controlled gate in (21.6) is the same as her applying the nonlocal controlled gate $\sum_m |m\rangle\langle m|^{A_1} \otimes \left(U^T(s(m))\right)^{B^n}$, due to the nonlocal (and perhaps spooky!) properties of the entangled state $|\varphi\rangle^{A^n B^n}$. Alice then sends her systems A^n through many uses of the noisy quantum channel $\mathcal{N}^{A\rightarrow B'}$, whose isometric extension is $U_{\mathcal{N}}^{A\rightarrow B'E}$. Let $|\varphi\rangle^{B'^n E^n B^n}$ denote the state resulting from the isometric extension $U_{\mathcal{N}}^{A\rightarrow B'E}$ of the channel acting on the state $|\varphi\rangle^{A^n B^n}$:

$$|\varphi\rangle^{B'^n E^n B^n} \equiv U_{\mathcal{N}}^{A^n \rightarrow B'^n E^n} |\varphi\rangle^{A^n B^n}. \tag{21.9}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

After Alice transmits through the channel, the state becomes

$$\sum_{l,m} \alpha_{l,m} |l\rangle^{R} |m\rangle^{A_1} \left(U^T(s(m))\right)^{B^n} |\varphi\rangle^{B'^n E^n B^n}, \tag{21.10}$$

where Bob now holds his shares B^n of the entanglement and the channel outputs B'^n. (Observe that the action of the controlled unitary in (21.6) commutes with the action of the channel.) Rather than perform an incoherent measurement with the POVM $\{\Lambda_m^{B'^n B^n}\}$, Bob applies a coherent gentle measurement (see Section 5.4), an isometry of the following form:

$$\sum_m \sqrt{\Lambda_m}^{B'^n B^n} \otimes |m\rangle^{B_1}. \tag{21.11}$$

Using the result of Exercise 5.4.1, we can readily check that the resulting state is $2\sqrt{\epsilon}$-close in trace distance to the following state:

$$\sum_{l,m} \alpha_{l,m} |l\rangle^{R} |m\rangle^{A_1} \left(U^T(s(m))\right)^{B^n} |\varphi\rangle^{B'^n E^n B^n} |m\rangle^{B_1}. \tag{21.12}$$

Thus, for the rest of the protocol, we pretend as if they are acting on the above state. Alice and Bob would like to coherently remove the coupling of their index m to the environment, so Bob performs the following controlled unitary:

$$\sum_m |m\rangle\langle m|^{B_1} \otimes (U^*(s(m)))^{B^n}, \tag{21.13}$$

and the final state is

$$\sum_{l,m=1}^{D^2} \alpha_{l,m} |l\rangle^{R} |m\rangle^{A_1} |\varphi\rangle^{B'^n E^n B^n} |m\rangle^{B_1}$$

$$= \left(\sum_{l,m=1}^{D^2} \alpha_{l,m} |l\rangle^{R} |m\rangle^{A_1} |m\rangle^{B_1} \right) \otimes |\varphi\rangle^{B'^n E^n B^n}. \tag{21.14}$$

Thus, this protocol implements a D^2-dimensional coherent channel up to an arbitrarily small error, and we have shown that the resource inequality in the statement of the theorem holds. Figure 21.1 depicts the entanglement-assisted coherent coding protocol. □

It is now a straightforward task to convert the protocol from Theorem 21.1.1 into one for entanglement-assisted quantum communication, by exploiting the coherent communication identity from Section 7.5.

Corollary 21.1.1 (Entanglement-Assisted Quantum Communication)**.** *The following resource inequality corresponds to an achievable protocol for entanglement-assisted quantum communication over a noisy quantum channel:*

$$\langle \mathcal{N} \rangle + \frac{1}{2} I(A;E)_{\varphi}[qq] \geq \frac{1}{2} I(A;B)_{\varphi}[q \to q], \tag{21.15}$$

where $|\varphi\rangle^{ABE} \equiv U_{\mathcal{N}}^{A' \to BE} |\varphi\rangle^{AA'}$ *and* $U_{\mathcal{N}}^{A' \to BE}$ *is an isometric extension of the channel* $\mathcal{N}^{A' \to B}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

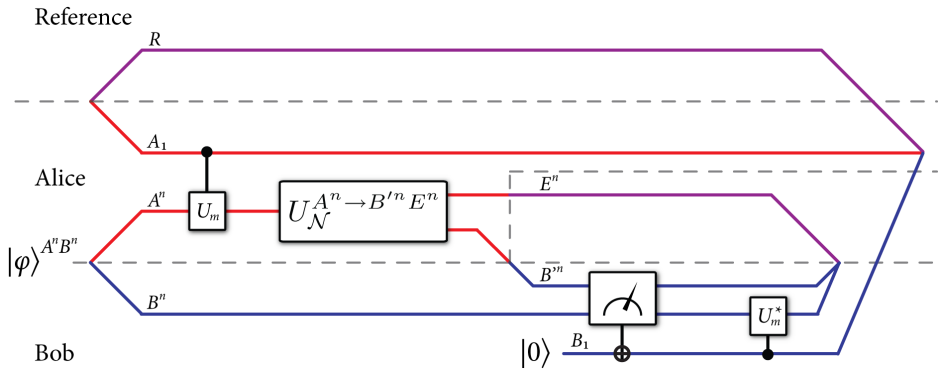

Figure 21.1: The protocol for entanglement-assisted coherent communication. Observe that it is the coherent version of the protocol for entanglement-assisted classical communication, just as coherent dense coding is the coherent version of super-dense coding (compare this figure and Figure 20.3 with Figures 6.2 and 7.3). Instead of applying conditional unitaries, Alice applies a controlled unitary from her system A_1 to her share of the entanglement and sends the encoded state through many uses of the noisy channel. Rather than performing a POVM, Bob performs a coherent gentle measurement from his systems B'^n and B^n to an ancilla B_1. Finally, he applies a similar controlled unitary in order to decouple the environment from the state of his ancilla B_1.

Consider the coherent communication identity from Section 7.5. This identity states that a D^2-dimensional coherent channel can perfectly simulate a D-dimensional quantum channel and a maximally entangled state $|\Phi\rangle^{AB}$ with Schmidt rank D. In terms of cobits, qubits, and ebits, the coherent communication identity is the following resource equality for D-dimensional systems:

$$2\log D[q \to qq] = \log D[q \to q] + \log D[qq]. \tag{21.16}$$

Consider the following chain of resource inequalities:

$$\langle \mathcal{N} \rangle + H(A)_\varphi[qq] \geq I(A;B)_\varphi[q \to qq] \tag{21.17}$$

$$\geq \frac{1}{2} I(A;B)_\varphi[q \to q] + \frac{1}{2} I(A;B)_\varphi[qq] \tag{21.18}$$

The first resource inequality is the statement of Theorem 21.1.1, and the second resource inequality follows from an application of coherent teleportation. If we then allow for catalytic protocols, in which we allow for some use of a resource with the demand that it be returned at the end of the protocol, we have a protocol for entanglement-assisted quantum communication:

$$\langle \mathcal{N} \rangle + \frac{1}{2} I(A;E)_\varphi[qq] \geq \frac{1}{2} I(A;B)_\varphi[q \to q], \tag{21.19}$$

because $H(A)_\varphi - \frac{1}{2} I(A;B)_\varphi = \frac{1}{2} I(A;E)_\varphi$ (see Exercise 11.6.6).

When comparing the entanglement consumption rate of the naive protocol in (21.1-21.2) with that of the protocol in Theorem 21.1.1, we see that the former requires an additional $I(A;B)_\rho$ ebits per channel use. Also, Theorem 21.1.1 leads to a simple proof of the achievability part of the quantum capacity theorem, as we see in the next section.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 21.1.1 Suppose that Alice can obtain the environment E of the channel $U_{\mathcal{N}}^{A'\to BE}$. Such a channel is known as a *coherent feedback isometry*. Show how they can achieve the following resource inequality with the coherent feedback isometry $U_{\mathcal{N}}^{A'\to BE}$:

$$\langle U_{\mathcal{N}}^{A'\to BE}\rangle \geq \frac{1}{2}I(A;B)_{\varphi}[q\to q] + \frac{1}{2}I(E;B)_{\varphi}[qq], \tag{21.20}$$

where $|\varphi\rangle^{ABE} = U_{\mathcal{N}}^{A'\to BE}|\varphi\rangle^{AA'}$ and $\rho^{A'} = \mathrm{Tr}_A\{\varphi^{AA'}\}$. This protocol is a generalization of coherent teleportation from Section 7.4 because it reduces to coherent teleportation in the case that $U_{\mathcal{N}}^{A'\to BE}$ is equivalent to two coherent channels.

21.2 Quantum Communication

We can obtain a protocol for quantum communication simply by combining the protocol from Theorem 21.1.1 further with entanglement distribution. The resulting protocol again makes catalytic use of entanglement, in the sense that it exploits some amount of entanglement shared between Alice and Bob at the beginning of the protocol, but it generates the same amount of entanglement at the end, so that the net entanglement consumption rate of the protocol is zero. The resulting rate of quantum communication turns out to be the same as we find for the quantum channel coding theorem in Chapter 23 (though the protocol given there does not make catalytic use of shared entanglement).

Corollary 21.2.1 (Quantum Communication)**.** *The coherent information $Q(\mathcal{N})$ is an achievable rate for quantum communication over a quantum channel $\mathcal{N}$. That is, the following resource inequality holds*

$$\langle\mathcal{N}\rangle \geq Q(\mathcal{N})[q\to q], \tag{21.21}$$

where $Q(\mathcal{N}) \equiv \max_{\varphi} I(A\rangle B)_{\rho}$ and $\rho^{AB} \equiv \mathcal{N}^{A'\to B}(\varphi^{AA'})$.

Proof. If we further combine the entanglement-assisted quantum communication protocol from Theorem 21.1.1 with entanglement distribution at a rate $\frac{1}{2}I(A;E)_{\rho}$, we obtain the following resource inequalities:

$$\begin{aligned}
&\langle\mathcal{N}\rangle + \frac{1}{2}I(A;E)_{\rho}[qq] \\
&\geq \frac{1}{2}\Big[I(A;B)_{\rho} - I(A;E)_{\rho}\Big][q\to q] + \frac{1}{2}I(A;E)_{\rho}[q\to q] &(21.22)\\
&\geq \frac{1}{2}\Big[I(A;B)_{\rho} - I(A;E)_{\rho}\Big][q\to q] + \frac{1}{2}I(A;E)_{\rho}[qq], &(21.23)
\end{aligned}$$

which after resource cancelation, becomes

$$\langle\mathcal{N}\rangle \geq I(A\rangle B)_{\rho}[q\to q], \tag{21.24}$$

because $I(A\rangle B)_{\rho} = \frac{1}{2}\Big[I(A;B)_{\rho} - I(A;E)_{\rho}\Big]$ (see Exercise 11.6.6). They can achieve the coherent information of the channel simply by generating codes from the state $\varphi^{AA'}$ that maximizes the channel's coherent information. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

21.3 Noisy Super-Dense Coding

Recall that the resource inequality for super-dense coding is

$$[q \to q] + [qq] \geq 2[c \to c]. \tag{21.25}$$

The entanglement-assisted classical communication protocol from the previous chapter is one way to generalize this protocol to a noisy setting, simply by replacing the noiseless qubit channels in (21.25) with many uses of a noisy quantum channel. This replacement leads to the setting of entanglement-assisted classical communication presented in the previous chapter.

Another way to generalize super-dense coding is to let the entanglement be noisy while keeping the quantum channels noiseless. We allow Alice and Bob access to many copies of some shared noisy state ρ^{AB} and to many uses of a noiseless qubit channel with the goal of generating noiseless classical communication. One might expect the resulting protocol to be similar to that for entanglement-assisted classical communication, and this is indeed the case. The resulting protocol is known as *noisy super-dense coding*:

Theorem 21.3.1 (Noisy Super-Dense Coding)**.** *The following resource inequality corresponds to an achievable protocol for quantum-assisted classical communication with a noisy quantum state:*

$$\langle \rho^{AB} \rangle + H(A)_\rho [q \to q] \geq I(A;B)_\rho [c \to c], \tag{21.26}$$

where ρ^{AB} is some noisy bipartite state that Alice and Bob share at the beginning of the protocol.

Proof. The proof of the existence of a protocol proceeds similarly to the proof of Theorem 20.4.1, with a few modifications to account for our different setting here. We simply need to establish a way for Alice and Bob to select a code randomly, and then we can invoke the Packing Lemma (Lemma 15.3.1) to establish the existence of a detection POVM that Bob can employ to detect Alice's messages. The method by which they select a random code is exactly the same as they do in the proof of Theorem 20.4.1, and for this reason, we only highlight the key aspects of the proof. First consider the state ρ^{AB}, and suppose that $|\varphi\rangle^{ABR}$ is a purification of this state, with R a reference system to which Alice and Bob do not have access. We can say that the state $|\varphi\rangle^{ABR}$ arises from some isometry $U_{\mathcal{N}}^{A' \to BR}$ acting on system A' of a pure state $|\varphi\rangle^{AA'}$, so that $|\varphi\rangle^{AA'}$ is defined by $|\varphi\rangle^{ABR} = U_{\mathcal{N}}^{A' \to BR} |\varphi\rangle^{AA'}$. We can also then think that the state ρ^{AB} arises from sending the state $|\varphi\rangle^{AA'}$ through a channel $\mathcal{N}^{A' \to B}$, obtained by tracing out the environment R of $U_{\mathcal{N}}^{A' \to BR}$. Our setting here is becoming closer to the setting in the proof of Theorem 20.4.1, and we now show how it becomes nearly identical. Observe that the state $(|\varphi\rangle^{AA'})^{\otimes n}$ admits a type decomposition, similar to the type decomposition in (20.40-20.43):

$$(|\varphi\rangle^{AA'})^{\otimes n} = \sum_t \sqrt{p(t)} |\Phi_t\rangle^{A^n A'^n}. \tag{21.27}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Similarly, we can write $\left(|\varphi\rangle^{ABR}\right)^{\otimes n}$ as

$$(|\varphi\rangle^{ABR})^{\otimes n} = \sum_t \sqrt{p(t)}|\Phi_t\rangle^{A^n \mid B^n R^n}, \tag{21.28}$$

where the vertical line in $A^n \mid B^n R^n$ indicates the bipartite cut between systems A^n and $B^n R^n$. Alice can select a unitary $U(s)^{A^n}$ of the form in (20.46) uniformly at random, and the expected density operator with respect to this random choice of unitary is

$$\overline{\rho}^{A^n B^n} \equiv \mathbb{E}_S\left\{U(S)^{A^n}\rho^{A^n B^n}U^\dagger(S)^{A^n}\right\} \tag{21.29}$$

$$= \sum_t p(t)\pi_t^{A^n} \otimes \mathcal{N}^{A'^n \to B^n}(\pi_t^{A'^n}), \tag{21.30}$$

by exploiting the development in (20.82-20.92). For each message m that Alice would like to send, she selects a vector s of the form in (20.47) uniformly at random, and we can write $s(m)$ to denote the explicit association of the vector s with the message m after Alice makes the assignment. This leads to quantum-assisted codewords[3] of the following form:

$$U(s(m))^{A^n}\rho^{A^n B^n}U^\dagger(s(m))^{A^n}. \tag{21.31}$$

We would now like to exploit the Packing Lemma (Lemma 15.3.1), and we require message subspace projectors and a total subspace projector in order to do so. We choose them respectively as

$$U(s)^{A^n}\Pi_{\rho,\delta}^{A^n B^n}U^\dagger(s)^{A^n}, \tag{21.32}$$

$$\Pi_{\rho,\delta}^{A^n} \otimes \Pi_{\rho,\delta}^{B^n}, \tag{21.33}$$

where $\Pi_{\rho,\delta}^{A^n B^n}$, $\Pi_{\rho,\delta}^{A^n}$, and $\Pi_{\rho,\delta}^{B^n}$ are typical projectors for $\rho^{A^n B^n}$, ρ^{A^n}, and ρ^{B^n}, respectively. The following four conditions for the Packing Lemma hold, for the same reasons that they hold in (20.68-20.71):

$$\operatorname{Tr}\left\{\left(\Pi_{\rho,\delta}^{A^n} \otimes \Pi_{\rho,\delta}^{B^n}\right)\left(U(s)^{A^n}\rho^{A^n B^n}U^\dagger(s)^{A^n}\right)\right\} \geq 1-\epsilon, \tag{21.34}$$

$$\operatorname{Tr}\left\{\left(U(s)^{A^n}\Pi_{\rho,\delta}^{A^n B^n}U^\dagger(s)^{A^n}\right)\left(U(s)^{A^n}\rho^{A^n B^n}U^\dagger(s)^{A^n}\right)\right\} \geq 1-\epsilon, \tag{21.35}$$

$$\operatorname{Tr}\left\{U(s)^{A^n}\Pi_{\rho,\delta}^{A^n B^n}U^\dagger(s)^{A^n}\right\} \leq 2^{n\left[H(AB)_\rho + c\delta\right]}, \tag{21.36}$$

$$\left(\Pi_{\rho,\delta}^{A^n} \otimes \Pi_{\rho,\delta}^{B^n}\right)\overline{\rho}^{A^n B^n}\left(\Pi_{\rho,\delta}^{A^n} \otimes \Pi_{\rho,\delta}^{B^n}\right)$$

$$\leq 2^{-n\left[H(A)_\rho + H(B)_\rho - \eta(n,\delta) - c\delta\right]}\left(\Pi_{\rho,\delta}^{A^n} \otimes \Pi_{\rho,\delta}^{B^n}\right), \tag{21.37}$$

[3]We say that the codewords are "quantum-assisted" because we will allow the assistance of quantum communication in transmitting them to Bob.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

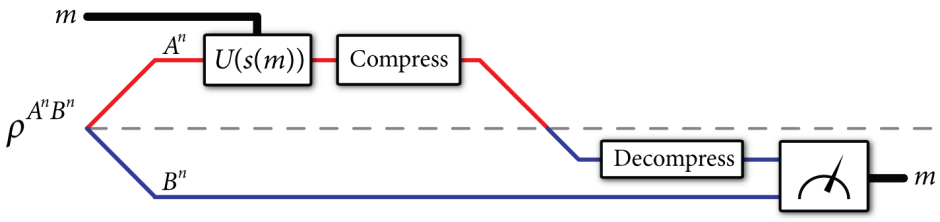

Figure 21.2: The protocol for noisy super-dense coding that corresponds to the resource inequality in Theorem 21.3.1. Alice first projects her share into its typical subspace (not depicted). She then applies a unitary encoding $U(s(m))$, based on her message m, to her share of the state $\rho^{A^nB^n}$. She compresses her state to approximately $nH(A)_\rho$ qubits and transmits these qubits over noiseless qubit channels. Bob decompresses the state and performs a decoding POVM that gives Alice's message m with high probability.

where c is some positive constant and $\eta(n,\delta)$ is a function that approaches zero as $n \to \infty$ and $\delta \to 0$. Let us assume for the moment that Alice simply sends her A^n systems to Bob with many uses of a noiseless qubit channel. It then follows from Corollary 15.5.1 (the derandomized version of the Packing Lemma) that there exists a code and a POVM $\{\Lambda_m^{A^nB^n}\}$ that can detect the transmitted codewords of the form in (21.31) with arbitrarily low maximal probability of error, as long as the size $|\mathcal{M}|$ of Alice's message set is small enough:

$$p_e^* \equiv \max_m \text{Tr}\left\{\left(I - \Lambda_m^{A^nB^n}\right)U(s(m))^{B^n}\rho^{A^nB^n}U^*(s(m))^{B^n}\right\} \tag{21.38}$$

$$\leq 4\left(\epsilon + 2\sqrt{\epsilon}\right) + 16 \cdot 2^{-n\left[H(A)_\rho + H(B)_\rho - \eta(n,\delta) - c\delta\right]}2^{n\left[H(AB)_\rho + c\delta\right]}|\mathcal{M}| \tag{21.39}$$

$$= 4\left(\epsilon + 2\sqrt{\epsilon}\right) + 16 \cdot 2^{-n\left[I(A;B)_\rho - \eta(n,\delta) - 2c\delta\right]}|\mathcal{M}|. \tag{21.40}$$

So, we can choose the size of the message set to be $|\mathcal{M}| = 2^{n[I(A;B)-\eta(n,\delta)-3c\delta]}$ so that the rate C of classical communication is

$$C = \frac{1}{n}\log_2|\mathcal{M}| = I(A;B)_\rho - \eta(n,\delta) - 3c\delta, \tag{21.41}$$

and the bound on the maximal probability of error becomes

$$p_e^* \leq 4\left(\epsilon + 2\sqrt{\epsilon}\right) + 16 \cdot 2^{-nc\delta}. \tag{21.42}$$

Since ϵ is an arbitrary positive number that approaches zero for sufficiently large n and δ is a positive constant, the maximal probability of error vanishes as n becomes large. Thus, the quantum mutual information $I(A;B)_\rho$, with respect to the state ρ^{AB} is an achievable rate for noisy super-dense coding with ρ. We now summarize the protocol (with a final modification). Alice and Bob begin with the state $\rho^{A^nB^n}$. Alice first performs a typical subspace measurement of her system A^n. This measurement succeeds with high probability and reduces the size of her system A^n to a subspace with size approximately equal to $nH(A)_\rho$ qubits. If Alice wishes to send message m, she applies the unitary $U(s(m))^{A^n}$ to her share

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

of the state. She then performs a compression isometry from her subspace of A^n to $nH(A)_\rho$ qubits. She transmits her qubits over $nH(A)_\rho$ noiseless qubit channels, and Bob receives them. Bob performs the decompression isometry from the space of $nH(A)_\rho$ noiseless qubits to a space isomorphic to Alice's original systems A^n. He then performs the decoding POVM $\{\Lambda_m^{A^nB^n}\}$ and determines Alice's message m with vanishingly small error probability. *Note*: The only modification to the protocol is the typical subspace measurement at the beginning, and one can readily check that this measurement does not affect any of the conditions in (21.34-21.37). Figure 21.2 depicts the protocol. □

21.4 State Transfer

We can also construct a coherent version of the noisy super-dense coding protocol, in a manner similar to the way in which the proof of Theorem 21.1.1 constructs a coherent version of entanglement-assisted classical communication. Though, the coherent version of noisy super-dense coding achieves an additional task: the transfer of Alice's share of the state $(\rho^{AB})^{\otimes n}$ to Bob. The resulting protocol is known as coherent state transfer, and from this protocol, we can derive a protocol for quantum-communication-assisted state transfer, or quantum-assisted state transfer[4] for short.

Theorem 21.4.1 (Coherent State Transfer)**.** *The following resource inequality corresponds to an achievable protocol for coherent state transfer with a noisy state ρ^{AB}:*

$$\langle W^{S\to AB} : \rho^S \rangle + H(A)_\rho[q \to q] \geq I(A;B)_\rho[q \to qq] + \langle I^{S\to \hat{B}B} : \rho^S \rangle, \tag{21.43}$$

where ρ^{AB} is some noisy bipartite state that Alice and Bob share at the beginning of the protocol.

The resource inequality in (21.43) features some notation that we have not seen yet. The expression $\langle W^{S\to AB} : \rho^S \rangle$ means that a source party S distributes many copies of the state ρ^S to Alice and Bob, by applying some isometry $W^{S\to AB}$ to the state ρ^S. This resource is effectively equivalent to Alice and Bob sharing many copies of the state ρ^{AB}, a resource we expressed in Theorem 21.3.1 as $\langle \rho^{AB} \rangle$. The expression $\langle I^{S\to \hat{B}B} : \rho^S \rangle$ means that a source party applies the identity map to ρ^S and gives the full state to Bob. We can now state the meaning of the resource inequality in (21.43): Using n copies of the state ρ^{AB} and $nH(A)_\rho$ noiseless qubit channels, Alice can simulate $nI(A;B)_\rho$ noiseless coherent channels to Bob while at the same time transferring her share of the state $(\rho^{AB})^{\otimes n}$ to him.

Proof. The proof proceeds similarly to the proof of Theorem 21.1.1. Let $|\varphi\rangle^{ABR}$ be a purification of ρ^{AB}. Alice begins with a state that she shares with a reference system R_1, on which she would like to simulate coherent channels:

$$|\psi\rangle^{R_1A_1} \equiv \sum_{l,m=1}^{D^2} \alpha_{l,m} |l\rangle^{R_1} |m\rangle^{A_1}, \tag{21.44}$$

[4]This protocol goes by several other names in the quantum Shannon theory literature: state transfer, fully-quantum Slepian-Wolf, state merging, and the merging mother.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $D^2 \approx 2^{nI(A;B)_\rho}$. She appends $|\psi\rangle^{R_1 A_1}$ to $|\varphi\rangle^{A^n B^n R^n} \equiv (|\varphi\rangle^{ABR})^{\otimes n}$ and applies a typical subspace measurement to her system A^n. (In what follows, we use the same notation for the typical projected state because the states are the same up to a vanishingly small error). She applies the following controlled unitary to her systems $A_1 A^n$:

$$\sum_m |m\rangle\langle m|^{A_1} \otimes U(s(m))^{A^n}, \tag{21.45}$$

resulting in the overall state:

$$\sum_{l,m} \alpha_{l,m} |l\rangle^{R_1} |m\rangle^{A_1} U(s(m))^{A^n} |\varphi\rangle^{A^n B^n R^n}. \tag{21.46}$$

Alice compresses her A^n systems, sends them over $nH(A)_\rho$ noiseless qubit channels, and Bob receives them. He decompresses them and places them in systems $\hat{B}^n$ isomorphic to A^n. The resulting state is the same as $|\varphi\rangle^{A^n B^n R^n}$, with the systems A^n replaced by $\hat{B}^n$. Bob performs a coherent gentle measurement of the following form:

$$\sum_m \sqrt{\Lambda_m}^{\hat{B}^n B^n} \otimes |m\rangle^{B_1}, \tag{21.47}$$

resulting in a state that is close in trace distance to

$$\sum_{l,m} \alpha_{l,m} |l\rangle^{R_1} |m\rangle^{A_1} |m\rangle^{B_1} U(s(m))^{\hat{B}^n} |\varphi\rangle^{\hat{B}^n B^n R^n}. \tag{21.48}$$

He finally performs the controlled unitary

$$\sum_m |m\rangle\langle m|^{B_1} \otimes U^\dagger(s(m))^{\hat{B}^n}, \tag{21.49}$$

resulting in the state

$$\left(\sum_{l,m} \alpha_{l,m} |l\rangle^{R_1} |m\rangle^{A_1} |m\rangle^{B_1} \right) \otimes |\varphi\rangle^{\hat{B}^n B^n R^n}. \tag{21.50}$$

Thus, Alice has simulated $nI(A;B)_\rho$ coherent channels to Bob with arbitrarily small error, while also transferring her share of the state $|\varphi\rangle^{A^n B^n R^n}$ to him. Figure 21.3 depicts the protocol. □

We obtain the following resource inequality for quantum-assisted state transfer, by combining the above protocol with the coherent communication identity:

Corollary 21.4.1 (Quantum-Assisted State Transfer)**.** *The following resource inequality corresponds to an achievable protocol for quantum-assisted state transfer with a noisy state ρ^{AB}:*

$$\langle W^{S\to AB} : \rho^S \rangle + \frac{1}{2} I(A;R)_\varphi [q \to q] \geq \frac{1}{2} I(A;B)_\varphi [qq] + \langle I^{S\to \hat{B}B} : \rho^S \rangle, \tag{21.51}$$

where ρ^{AB} is some noisy bipartite state that Alice and Bob share at the beginning of the protocol, and $|\varphi\rangle^{ABR}$ is a purification of it.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

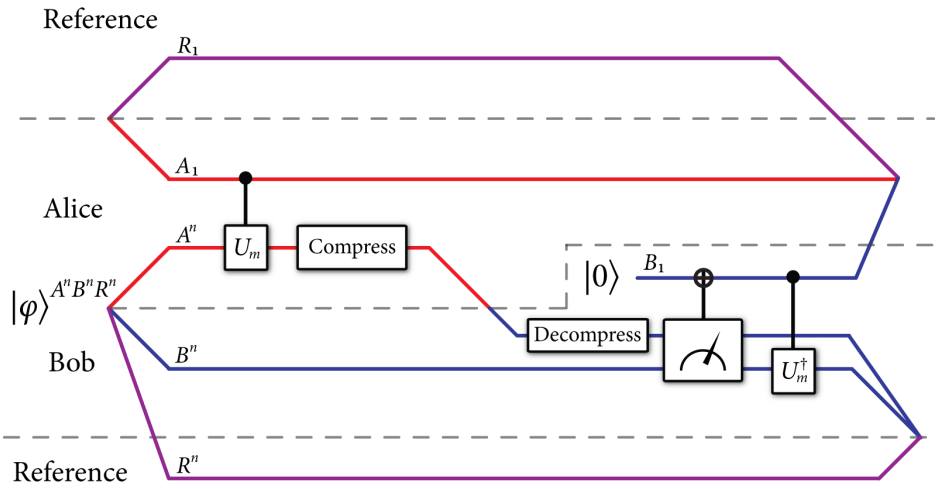

Figure 21.3: The protocol for coherent state transfer, a coherent version of the noisy super-dense coding protocol that accomplishes the task of state transfer in addition to coherent communication.

Proof. Consider the following chain of resource inequalities:

$$\begin{aligned}\langle W^{S\to AB} : \rho^S\rangle + H(A)_\varphi[qq] & \\ \geq I(A;B)_\varphi[q\to qq] + \langle I^{S\to\hat{B}B} : \rho^S\rangle & \quad (21.52)\\ \geq \frac{1}{2}I(A;B)_\varphi[q\to q] + \frac{1}{2}I(A;B)_\varphi[qq] + \langle I^{S\to\hat{B}B} : \rho^S\rangle, & \quad (21.53)\end{aligned}$$

where the first follows from coherent state transfer and the second follows from the coherent communication identity. By resource cancelation, we obtain the resource inequality in the statement of the theorem because $\frac{1}{2}I(A;R)_\varphi = H(A)_\rho - \frac{1}{2}I(A;B)_\rho$. □

Corollary 21.4.2 (Classical-Assisted State Transfer)**.** *The following resource inequality corresponds to an achievable protocol for classical-assisted state transfer with a noisy state ρ^{AB}:*

$$\langle W^{S\to AB} : \rho^S\rangle + I(A;R)_\varphi[c\to c] \geq I(A\rangle B)_\varphi[qq] + \langle I^{S\to\hat{B}B} : \rho^S\rangle, \quad (21.54)$$

where ρ^{AB} is some noisy bipartite state that Alice and Bob share at the beginning of the protocol, and $|\varphi\rangle^{ABR}$ is a purification of it.

Proof. We simply combine the protocol above with teleportation:

$$\begin{aligned}\langle W^{S\to AB} : \rho^S\rangle + \frac{1}{2}I(A;R)_\varphi[q\to q] + I(A;R)_\varphi[c\to c] + \frac{1}{2}I(A;R)_\varphi[qq] & \\ \geq \frac{1}{2}I(A;B)_\varphi[qq] + \langle I^{S\to\hat{B}B} : \rho^S\rangle + \frac{1}{2}I(A;R)_\varphi[q\to q] & \quad (21.55)\end{aligned}$$

Canceling terms for both quantum communication and entanglement, we obtain the resource inequality in the statement of the corollary. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above protocol gives a wonderful operational interpretation to the coherent information (or negative conditional entropy $-H(A|B)_\rho$). When the coherent information is positive, Alice and Bob share that rate of entanglement at the end of the protocol (and thus the ability to teleport if extra classical communication is available). When the coherent information is negative, they need to consume entanglement at a rate of $H(A|B)_\rho$ ebits per copy in order for the state transfer process to complete.

Exercise 21.4.1 Suppose that Alice actually possesses the reference R in the above protocols. Show that Alice and Bob can achieve the following resource inequality:

$$\langle \psi^{ABR} \rangle + \frac{1}{2} I(A;R)_\psi [q \to q] \geq \frac{1}{2}\Big(H(A)_\psi + H(B)_\psi + H(R)_\psi \Big)[qq], \tag{21.56}$$

where $|\psi\rangle^{ABR}$ is some pure state.

21.4.1 The Dual Roles of Quantum Mutual Information

The resource inequality for entanglement-assisted quantum communication in (21.15) and that for quantum-assisted state transfer in (21.51) appear to be strikingly similar. Both contain a noisy resource and both consume a noiseless quantum resource in order to generate another noiseless quantum resource. We say that these two protocols are related by *source-channel duality* because we obtain one protocol from another by changing channels to states and vice versa.

Also, both protocols require the consumed rate of the noiseless quantum resource to be equal to half the quantum mutual information between the system A for which we are trying to preserve quantum coherence and the environment to which we do not have access. In both cases, our goal is to break the correlations between the system A and the environment, and the quantum mutual information is quantifying how much quantum coherence is required to break these correlations. Both protocols in (21.15) and (21.51) have their rates for the generated noiseless quantum resource equal to half the quantum mutual information between the system A and the system B. Thus, the quantum mutual information is also quantifying how much quantum correlations we can establish between two systems—it plays the dual role of quantifying both the destruction and creation of correlations.

21.5 Trade-off Coding

Suppose that you are a communication engineer working at a quantum communication company named *EA-USA*. Suppose further that your company has made quite a profit from entanglement-assisted classical communication, beating out the communication rates that other companies can achieve simply because your company has been able to generate high-quality noiseless entanglement between several nodes in its network, while the competitors have not been able to do so. But now suppose that your customer base has become so large that there is not enough entanglement to support protocols that achieve the rates given in

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

the entanglement-assisted classical capacity theorem (Theorem 20.3.1). Your boss would like you to make the best of this situation, by determining the optimal rates of classical communication for a fixed entanglement budget. He is hoping that you will be able to design a protocol such that there will only be a slight decrease in communication rates. You tell him that you will do your best.

What should you do in this situation? Your first thought might be that we have already determined unassisted classical codes with a communication rate equal to the channel Holevo information $\chi(\mathcal{N})$ and we have also determined entanglement-assisted codes with a communication rate equal to the channel mutual information $I(\mathcal{N})$. It might seem that a reasonable strategy is to mix these two strategies, using some fraction λ of the channel uses for the unassisted classical code and the other fraction $1-\lambda$ of the channel uses for the entanglement-assisted code. This strategy achieves a rate of

$$\lambda\ \chi(\mathcal{N}) + (1-\lambda) I(\mathcal{N}), \tag{21.57}$$

and it has an error no larger than the sum of the errors of the individual codes (thus, this error vanishes asymptotically). Meanwhile, it consumes entanglement at a lower rate of $(1-\lambda)E$ ebits per channel use, if E is the amount of entanglement that the original protocol for entanglement-assisted classical communication consumes. This simple mixing strategy is known as *time-sharing*. You figure this strategy might perform well, and you suggest it to your boss. After your boss reviews your proposal, he sends it back to you, telling you that he already thought of this solution and suggests that you are going to have to be a bit more clever—otherwise, he suspects that the existing customer base will notice the drop in communication rates.

Another strategy for communication is known as *trade-off coding*. We explore this strategy in the forthcoming section and in a broader context in Chapter 24. Trade-off coding beats time-sharing for many channels of interest, but for other channels, it just reduces to time-sharing. It is not clear *a priori* how to determine which channels benefit from trade-off coding, but it certainly depends on the channel for which Alice and Bob are coding. Chapter 24 follows up on the development here by demonstrating that this trade-off coding strategy is provably optimal for certain channels, and for general channels, it is optimal in the sense of regularized formulas. Trade-off coding is our best known way to deal with the above situation with a fixed entanglement budget, and your boss should be pleased with these results. Furthermore, we can upgrade the protocol outlined below to one that achieves entanglement-assisted communication of both classical and quantum information.

21.5.1 Trading between Unassisted and Assisted Classical Communication

We first show that the resource inequality given in the following theorem is achievable, and we follow up with an interpretation of it in the context of trade-off coding. We name the protocol *CE trade-off coding* because it captures the trade-off between classical communication and entanglement consumption.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Theorem 21.5.1 (CE Trade-off Coding)**.** *The following resource inequality corresponds to an achievable protocol for entanglement-assisted classical communication over a noisy quantum channel*

$$\langle \mathcal{N} \rangle + H(A|X)_\rho[qq] \geq I(AX;B)_\rho[c \to c], \tag{21.58}$$

where ρ^{XAB} *is a state of the following form:*

$$\rho^{XAB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A' \to B}(\varphi_x^{AA'}), \tag{21.59}$$

and the states $\varphi_x^{AA'}$ *are pure.*

Proof. The proof of the above trade-off coding theorem exploits the direct parts of both the HSW coding theorem (Theorem 19.3.1) and the entanglement-assisted classical capacity theorem (Theorem 20.4.1). In particular, we exploit the fact that the HSW codewords in Theorem 19.3.1 arise from strongly typical sequences and that the entanglement-assisted quantum codewords from Theorem 20.4.1 are tensor power states after tracing over Bob's shares of the entanglement (this is the observation in Remark 20.4.1). Suppose that Alice and Bob exploit an HSW code for the channel $\mathcal{N}^{A' \to B}$. Such a code consists of a codebook $\{\rho_{x^n(m)}\}_m$ with $\approx 2^{nI(X;B)_\rho}$ quantum codewords. The Holevo information $I(X;B)_\rho$ is with respect to some classical-quantum state ρ^{XB} where

$$\rho^{XB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A' \to B}(\rho_x^{A'}), \tag{21.60}$$

and each codeword $\rho_{x^n(m)}$ is a tensor-product state of the form

$$\rho_{x^n(m)} = \rho_{x_1(m)} \otimes \rho_{x_2(m)} \otimes \cdots \otimes \rho_{x_n(m)}. \tag{21.61}$$

Corresponding to the codebook is some decoding POVM $\{\Lambda_m^{B^n}\}$, which Bob can employ to decode each codeword transmitted through the channel with arbitrarily high probability for all $\epsilon > 0$:

$$\forall m \quad \mathrm{Tr}\Big\{\Lambda_m^{B^n} \mathcal{N}^{A'^n \to B^n}(\rho_{x^n(m)}^{A'^n})\Big\} \geq 1 - \epsilon. \tag{21.62}$$

Recall from the direct part of Theorem 19.3.1 that we select each codeword from the set of strongly typical sequences for the distribution $p_X(x)$ (see Definition 13.7.2). This implies that each classical codeword $x^n(m)$ has approximately $np_X(a_1)$ occurrences of the symbol $a_1 \in \mathcal{X}$, $np_X(a_2)$ occurrences of the symbol $a_2 \in \mathcal{X}$, and so on, for all letters in the alphabet $\mathcal{X}$. Without loss of generality and for simplicity, we assume that each codeword $x^n(m)$ has exactly these numbers of occurrences of the symbols in the alphabet $\mathcal{X}$. Then for any strongly typical sequence x^n, there exists some permutation π that arranges it in lexicographical order according to the alphabet $\mathcal{X}$. That is, this permutation arranges the sequence x^n into $|\mathcal{X}|$ blocks, each of length $np_X(a_1), \ldots, np_X(a_{|\mathcal{X}|})$:

$$\pi(x^n) = \underbrace{a_1 \cdots a_1}_{np_X(a_1)} \underbrace{a_2 \cdots a_2}_{np_X(a_2)} \cdots \underbrace{a_{|\mathcal{X}|} \cdots a_{|\mathcal{X}|}}_{np_X(a_{|\mathcal{X}|})}. \tag{21.63}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The same holds true for the corresponding permutation operator π applied to a quantum state ρ_{x^n} generated from a strongly typical sequence x^n:

$$\pi(\rho_{x^n}) = \underbrace{\rho_{a_1} \otimes \cdots \otimes \rho_{a_1}}_{np_X(a_1)} \otimes \underbrace{\rho_{a_2} \otimes \cdots \otimes \rho_{a_2}}_{np_X(a_2)} \otimes \cdots \otimes \underbrace{\rho_{a_{|\mathcal{X}|}} \otimes \cdots \otimes \rho_{a_{|\mathcal{X}|}}}_{np_X\left(a_{|\mathcal{X}|}\right)}. \tag{21.64}$$

Now, we assume that n is quite large, so large that each of $np_X(a_1)$, $\ldots$, $np_X(a_{|\mathcal{X}|})$ are large enough for the law of large numbers to come into play for each block in the permuted sequence $\pi(x^n)$ and tensor-product state $\pi(\rho_{x^n})$. Let $\varphi_x^{AA'}$ be a purification of each $\rho_x^{A'}$ in the ensemble $\left\{p_X(x), \rho_x^{A'}\right\}$, where we assume that Alice has access to system A' and Bob has access to A. Then, for every HSW quantum codeword $\rho_{x^n(m)}^{A'}$, there is some purification $\varphi_{x^n(m)}^{A^nA'^n}$, where

$$\varphi_{x^n(m)}^{A^nA'^n} \equiv \varphi_{x_1(m)}^{A_1A'_1} \otimes \varphi_{x_2(m)}^{A_2A'_2} \otimes \cdots \otimes \varphi_{x_n(m)}^{A_nA'_n}, \tag{21.65}$$

Alice has access to the systems $A'^n \equiv A'_1 \cdots A'_n$, and Bob has access to $A^n \equiv A_1 \cdots A_n$. Applying the permutation π to any purified tensor-product state φ_{x^n} gives

$$\pi(\varphi_{x^n}) = \underbrace{\varphi_{a_1} \otimes \cdots \otimes \varphi_{a_1}}_{np_X(a_1)} \otimes \underbrace{\varphi_{a_2} \otimes \cdots \otimes \varphi_{a_2}}_{np_X(a_2)} \otimes \cdots \otimes \underbrace{\varphi_{a_{|\mathcal{X}|}} \otimes \cdots \otimes \varphi_{a_{|\mathcal{X}|}}}_{np_X\left(a_{|\mathcal{X}|}\right)}, \tag{21.66}$$

where we have assumed that the permutation applies on both the purification systems A^n and the systems A'^n. We can now formulate a strategy for trade-off coding. Alice begins with a standard classical sequence $\hat{x}^n$ that is in lexicographical order, having exactly $np_X(a_i)$ occurrences of the symbol $a_i \in \mathcal{X}$ (of the form in (21.63)). According to this sequence, she arranges the states $\{\varphi_{a_i}^{AA'}\}$ to be in $|\mathcal{X}|$ blocks, each of length $np_X(a_i)$—the resulting state is of the same form as in (21.66). Since $np_X(a_i)$ is large enough for the law of large numbers to come into play, for each block, there exists an entanglement-assisted classical code with $\approx 2^{nI(A;B)_{\mathcal{N}(\varphi_{a_i})}}$ entanglement-assisted quantum codewords, where the quantum mutual information $I(A;B)_{\mathcal{N}(\varphi_{a_i})}$ is with respect to the state $\mathcal{N}^{A'\to B}(\varphi_{a_i}^{AA'})$. Let $n_i \equiv np_X(a_i)$. Then each of these $|\mathcal{X}|$ entanglement-assisted classical codes consumes $n_iH(A)_{\varphi_{a_i}^A}$ ebits. The entanglement-assisted quantum codewords for each block are of the form:

$$U(s(l_i))^{A^{n_i}}(\varphi_{a_i}^{A^{n_i}A'^{n_i}})U^\dagger(s(l_i))^{A^{n_i}}, \tag{21.67}$$

where l_i is a message in the message set of size $\approx 2^{nI(A;B)_{\varphi_{a_i}}}$, the state $\varphi_{a_i}^{A^{n_i}A'^{n_i}} = \varphi_{a_i}^{A_1A'_1} \otimes \cdots \otimes \varphi_{a_i}^{A_{n_i}A'_{n_i}}$, and the unitaries $U(s(l_i))^{A^{n_i}}$ are of the form in (20.46). Observe that the codewords in (21.67) are all equal to $\rho_{a_i}^{A'^{n_i}}$ after tracing over Bob's systems A^{n_i}, regardless of the particular unitary that Alice applies (this is the content of Remark 20.4.1). Alice then determines the permutation π_m needed to permute the standard sequence $\hat{x}^n$ to a codeword sequence $x^n(m)$, and she applies the permutation operator π_m to her systems A'^n so that her channel input density operator is the HSW quantum codeword $\rho_{x^n(m)}^{A'^n}$ (we are tracing over Bob's systems A^n and applying Remark 20.4.1 to obtain this result). She transmits

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

her systems A'^n over the channel to Bob. If Bob ignores his share of the entanglement in A^n, the state that he receives from the channel is $\mathcal{N}^{A'^n \to B^n}(\rho_{x^n(m)}^{A'^n})$. He then applies his HSW measurement $\{\Lambda_m^{B^n}\}$ to the systems B^n received from the channel, and he determines the sequence $x^n(m)$, and hence the message m, with nearly unit probability. Also, this measurement has negligible disturbance on the state, so that the post-measurement state is $2\sqrt{\epsilon}$-close in trace distance to the state that Alice transmitted through the channel (in what follows, we assume that the measurement does not change the state, and we collect error terms at the end of the proof). Now that he knows m, he applies the inverse permutation operator π_m^{-1} to his systems B^n, and we are assuming that he already has his share A^n of the entanglement arranged in lexicographical order according to the standard sequence $\hat{x}^n$. His state is then as follows:

$$\bigotimes_{i=1}^{|\mathcal{X}|} U(s(l_i))^{A^{n_i}} \left(\varphi_{a_1}^{A^{n_1}A'^{n_1}}\right) U^\dagger(s(l_i))^{A^{n_i}}. \tag{21.68}$$

At this point, he can decode the message l_i in the i^{th} block by performing a collective measurement on the systems $A^{n_i}A'^{n_i}$. He does this for each of the $|\mathcal{X}|$ entanglement-assisted classical codes, and this completes the protocol for trade-off coding. The total error accumulated in this protocol is no larger than the sum of ϵ for the first measurement, $2\sqrt{\epsilon}$ for the disturbance of the state, and $|\mathcal{X}|\epsilon$ for the error from the final measurement of the $|\mathcal{X}|$ blocks. The proof here assumes that every classical codeword $x^n(m)$ has exactly $np_X(a_i)$ occurrences of symbol $a_i \in \mathcal{X}$, but it is straightforward to modify the above protocol to allow for imprecision, i.e., if the codewords are δ-strongly typical. Figure 21.4 depicts this protocol for an example. We now show how the total rate of classical communication adds up to $I(AX;B)_\rho$ where ρ^{XAB} is a state of the form in (21.59). First, we can apply the chain rule for quantum mutual information to observe that the total rate $I(AX;B)_\rho$ is the sum of a Holevo information $I(X;B)_\rho$ and a classically conditioned quantum mutual information $I(A;B|X)_\rho$:

$$I(AX;B)_\rho = I(X;B)_\rho + I(A;B|X)_\rho. \tag{21.69}$$

They achieve the rate $I(X;B)_\rho$ because Bob first reliably decodes the HSW quantum codeword, of which there can be $\approx 2^{nI(X;B)}$. His next step is to permute and decode the $|\mathcal{X}|$ blocks, each consisting of an entanglement-assisted classical code on $\approx np_X(x)$ channel uses. Each entanglement-assisted classical code can communicate $np_X(x)I(A;B)_{\rho_x}$ bits while consuming $np_X(x)H(A)$ ebits. Thus, the total rate of classical communication for this last part is

$$\frac{\text{\# of bits generated}}{\text{\# of channel uses}} \approx \frac{\sum_x n\ p_X(x) I(A;B)_{\rho_x}}{\sum_x n\ p_X(x)} \tag{21.70}$$

$$= \sum_x p_X(x) I(A;B)_{\rho_x} \tag{21.71}$$

$$= I(A;B|X)_\rho. \tag{21.72}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

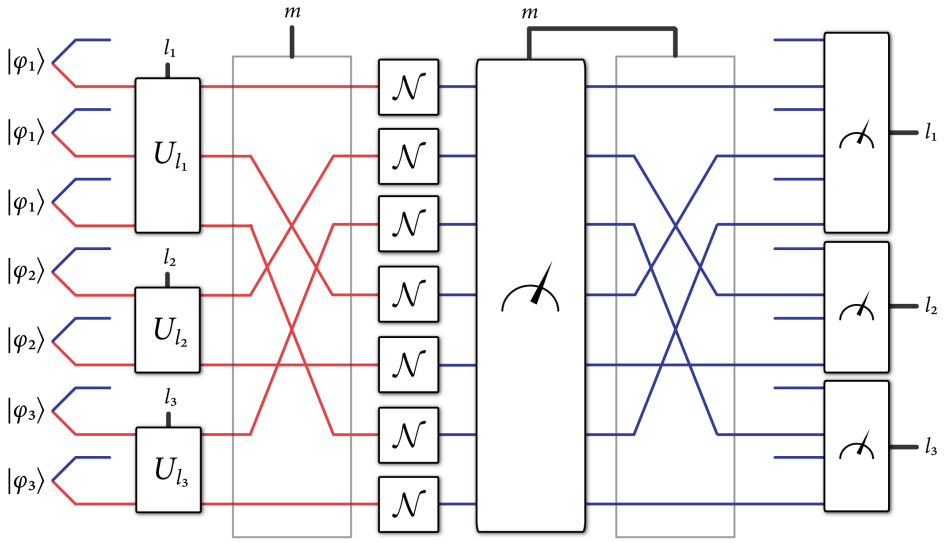

Figure 21.4: A simple protocol for trade-off coding between assisted and unassisted classical communication. Blue systems belong to Bob, and red systems belong to Alice. Alice wishes to send the classical message m while also sending the messages l_1, l_2, and l_3. Her HSW codebook has the message m map to the sequence 1231213, which in turn gives the HSW quantum codeword $\rho_1 \otimes \rho_2 \otimes \rho_3 \otimes \rho_1 \otimes \rho_2 \otimes \rho_1 \otimes \rho_3$. A purification of these states is the following tensor product of pure states: $\varphi_1 \otimes \varphi_2 \otimes \varphi_3 \otimes \varphi_1 \otimes \varphi_2 \otimes \varphi_1 \otimes \varphi_3$, where Bob possesses the purification of each state in the tensor product. She begins with these states arranged in lexicographic order in three blocks (there are three letters in this alphabet). For each block i, she encodes the message l_i with the local unitaries for an entanglement-assisted classical code. She then permutes her shares of the entangled states according to the permutation associated with the message m. She inputs her systems to many uses of the channel, and Bob receives the outputs. His first action is to ignore his shares of the entanglement and perform a collective HSW measurement on all of the channel outputs. With high probability, he can determine the message m while causing a negligible disturbance to the state of the channel outputs. Based on the message m, he performs the inverse of the permutation that Alice used at the encoder. He combines his shares of the entanglement with the permuted channel outputs. His final three measurements are those given by the three entanglement-assisted codes Alice used at the encoder, and they detect the messages l_1, l_2, and l_3 with high probability.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and similarly, the total rate of entanglement consumption is

$$\frac{\#\text{ of ebit consumed}}{\#\text{ of channel uses}} \approx \frac{\sum_x n\ p_X(x) H(A)_{\rho_x}}{\sum_x n\ p_X(x)} \tag{21.73}$$

$$= \sum_x p_X(x) H(A)_{\rho_x} \tag{21.74}$$

$$= H(A|X)_\rho. \tag{21.75}$$

This gives the resource inequality in the statement of the theorem. □

21.5.2 Trade-off Coding Subsumes Time-Sharing

Before proceeding to other trade-off coding settings, we show how time-sharing emerges as a special case of a trade-off coding strategy. Recall from (21.57) that time-sharing can achieve the rate $\lambda\ \chi(\mathcal{N}) + (1-\lambda) I(\mathcal{N})$ for any λ such that $0 \leq \lambda \leq 1$. Suppose that $\phi^{AA'}$ is the pure state that maximizes the channel mutual information $I(\mathcal{N})$, and suppose that $\{p_X(x), \psi_x^{A'}\}$ is an ensemble of pure states that maximizes the channel Holevo information $\chi(\mathcal{N})$ (recall from Theorem 12.3.2 that it is sufficient to consider pure states for maximizing the Holevo information of a channel). Time-sharing simply mixes between these two strategies, and we can construct a classical-quantum state of the form in (21.59), for which time-sharing turns out to be the strategy executed by the constructed trade-off code:

$$\sigma^{UXAB} \equiv (1-\lambda)|0\rangle\langle 0|^U \otimes |0\rangle\langle 0|^X \otimes \mathcal{N}^{A'\to B}(\phi^{AA'}) + \lambda |1\rangle\langle 1|^U \otimes \sum_x p_X(x) |x\rangle\langle x|^X \otimes |0\rangle\langle 0|^A \otimes \mathcal{N}^{A'\to B}(\psi_x^{A'}). \tag{21.76}$$

In the above, the register U is acting as a classical binary flag to indicate whether the code should be an entanglement-assisted classical capacity achieving code or a code that achieves the channel's Holevo information. The amount of classical bits that Alice can communicate to Bob with a trade-off code is $I(AUX;B)_\sigma$, where we have assumed that U and X together form the classical register. We can then evaluate this mutual information by applying the chain rule:

$$I(AUX;B)_\sigma = I(A;B|XU)_\sigma + I(X;B|U)_\sigma + I(U;B)_\sigma \tag{21.77}$$

$$= (1-\lambda) I(A;B)_{\mathcal{N}(\phi)} + \lambda \left[\sum_x p_X(x) I(A;B)_{|0\rangle\langle 0| \otimes \mathcal{N}(\psi_x)} \right] + (1-\lambda) I(X;B)_{|0\rangle\langle 0| \otimes \mathcal{N}(\phi)} + \lambda I(X;B)_{\{p(x),\psi_x\}} + I(U;B)_\sigma \tag{21.78}$$

$$\geq (1-\lambda) I(\mathcal{N}) + \lambda\ \chi(\mathcal{N}). \tag{21.79}$$

The second equality follows by evaluating the first two conditional mutual informations. The inequality follows from the assumptions that $I(\mathcal{N}) = I(A;B)_{\mathcal{N}(\phi)}$ and $\chi(\mathcal{N}) = I(X;B)_{\{p(x),\psi_x\}}$, the fact that quantum mutual information vanishes on product states, and $I(U;B)_\sigma \geq 0$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Thus, in certain cases, this strategy might do slightly better than time-sharing, but for channels for which $\phi^{A'} = \sum_x p(x)\psi_x^{A'}$, this strategy is equivalent to time-sharing because $I(U;B)_\sigma = 0$ in this latter case.

Thus, time-sharing emerges as a special case of trade-off coding. In general, we can try to see if trade-off coding beats time-sharing for certain channels by optimizing the rates in Theorem 21.5.1 over all possible choices of states of the form in (21.59).

21.5.3 Trading between Coherent and Classical Communication

We obtain the following corollary of Theorem 21.5.1, simply by upgrading the $|\mathcal{X}|$ entanglement-assisted classical codes to entanglement-assisted coherent codes. The upgrading is along the same lines as that in the proof of Theorem 21.1.1, and for this reason, we omit the proof.

Corollary 21.5.1. *The following resource inequality corresponds to an achievable protocol for entanglement-assisted coherent communication over a noisy quantum channel $\mathcal{N}$:*

$$\langle \mathcal{N} \rangle + H(A|X)_\rho[qq] \geq I(A;B|X)_\rho[q \to qq] + I(X;B)_\rho[c \to c], \tag{21.80}$$

where ρ^{XAB} is a state of the following form:

$$\rho^{XAB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A'\to B}(\varphi_x^{AA'}), \tag{21.81}$$

and the states $\varphi_x^{AA'}$ are pure.

21.5.4 Trading between Classical Communication and Entanglement-Assisted Quantum Communication

We end this section with a protocol that achieves entanglement-assisted communication of both classical and quantum information. It is essential to the trade-off between a noisy quantum channel and the three resources of noiseless classical communication, noiseless quantum communication, and noiseless entanglement. We study this trade-off in full detail in Chapter 24, where we show that combining this protocol with teleportation, super-dense coding, and entanglement distribution is sufficient to achieve any task in dynamic quantum Shannon theory involving the three unit resources.

Corollary 21.5.2 (CQE Trade-off Coding)**.** *The following resource inequality corresponds to an achievable protocol for entanglement-assisted communication of classical and quantum information over a noisy quantum channel*

$$\langle \mathcal{N} \rangle + \frac{1}{2} I(A;E|X)_\rho[qq] \geq \frac{1}{2} I(A;B|X)_\rho[q \to q] + I(X;B)_\rho[c \to c], \tag{21.82}$$

where ρ^{XAB} is a state of the following form:

$$\rho^{XABE} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes U_{\mathcal{N}}^{A'\to BE}(\varphi_x^{AA'}), \tag{21.83}$$

the states $\varphi_x^{AA'}$ are pure, and $U_{\mathcal{N}}^{A'\to BE}$ is an isometric extension of the channel $\mathcal{N}^{A'\to B}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Proof. Consider the following chain of resource inequalities:

$$\begin{aligned}
&\langle \mathcal{N} \rangle + H(A|X)_\rho[qq] \\
&\geq I(A;B|X)_\rho[q \to qq] + I(X;B)_\rho[c \to c] && (21.84) \\
&\geq \frac{1}{2} I(A;B|X)_\rho[qq] + \frac{1}{2} I(A;B|X)_\rho[q \to q] + I(X;B)_\rho[c \to c]. && (21.85)
\end{aligned}$$

The first inequality is the statement in Corollary 21.5.1, and the second inequality follows from the coherent communication identity. After resource cancelation and noting that $H(A|X)_\rho - \frac{1}{2} I(A;B|X)_\rho = \frac{1}{2} I(A;E|X)_\rho$, the resulting resource inequality is equivalent to the one in (21.82). □

21.5.5 Trading between Classical and Quantum Communication

Our final trade-off coding protocol that we consider is that between classical and quantum communication. The proof of the below resource inequality follows by combining the protocol in Corollary 21.5.2 with entanglement distribution, in much the same way as we did in Corollary 21.2.1. Thus, we omit the proof.

Corollary 21.5.3 (CQ Trade-off Coding)**.** *The following resource inequality corresponds to an achievable protocol for simultaneous classical and quantum communication over a noisy quantum channel*

$$\langle \mathcal{N} \rangle \geq I(A\rangle BX)_\rho[q \to q] + I(X;B)_\rho[c \to c], \tag{21.86}$$

where ρ^{XAB} *is a state of the following form:*

$$\rho^{XAB} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes \mathcal{N}^{A' \to B}(\varphi_x^{AA'}), \tag{21.87}$$

and the states $\varphi_x^{AA'}$ *are pure.*

21.6 Concluding Remarks

The maintainence of quantum coherence is the theme of this chapter. Alice and Bob can execute powerful protocols if they perform encoding and decoding in superposition. In both entanglement-assisted coherent communication and coherent state transfer, Alice performs controlled gates instead of conditional gates and Bob performs coherent measurements that place measurement outcomes in an ancilla register without destroying superpositions. Also, Bob's final action in both of these protocols is to perform a controlled decoupling unitary, ensuring that the state of the environment is independent of Alice and Bob's final state. Thus, the same protocol accomplishes the different tasks of entanglement-assisted coherent communication and coherent state transfer, and these in turn can generate a whole host of other protocols by combining them with entanglement distribution and the coherent and incoherent versions of teleportation and super-dense coding. Among these other

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

generated protocols are entanglement-assisted quantum communication, quantum communication, quantum-assisted state transfer, and classical-assisted state transfer. The exercises in this chapter explore further possibilities if Alice has access to the environments of the different protocols—the most general version of coherent teleportation arises in such a case.

Trade-off coding is the theme of the last part of this chapter. Here, we are addressing the question: Given a fixed amount of a certain resource, how much of another resource can Alice and Bob generate? Noisy quantum channels are the most fundamental description of a medium over which information can propagate, and it is thus important to understand the best ways to make effective use of such a resource for a variety of purposes. We determined a protocol that achieves the task of entanglement-assisted communication of classical and quantum information, simply by combining the protocols we have already found for classical communication and entanglement-assisted coherent communication. Chapter 24 continues this theme of trade-off coding in a much broader context and demonstrates that the protocol given here, when combined with teleportation, super-dense coding, and entanglement distribution, is optimal for some channels of interest and essentially optimal in the general case.

21.7 History and Further Reading

Devetak *et al.* showed that it was possible to make the protocols for entanglement-assisted classical communication and noisy super-dense coding coherent [71, 70], leading to Theorems 21.1.1 and 21.4.1. They called these protocols the "father" and "mother," respectively, because they generated many other protocols in quantum Shannon theory by combining them with entanglement distribution, teleportation, and super-dense coding. Horodecki *et al.* [152] formulated a protocol for noisy super-dense coding, but our protocol here makes use of the coding technique in Ref. [156]. Shor first proved a coding theorem for trading between assisted and unassisted classical communication [229], and Devetak and Shor followed up on this result by finding a scheme for trade-off coding between classical and quantum communication. Some time later, Ref. [159] generalized these two coding schemes to produce the result of Theorem 21.5.2.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 22

Private Classical Communication

We have now seen in Chapters 19-21 how Alice can communicate classical or quantum information to Bob, perhaps even with the help of shared entanglement. One might argue that these communication tasks are the most fundamental tasks in quantum Shannon theory, given that they have furthered our understanding of the nature of information transmission over quantum channels. Though, when discussing the communication of classical information, we made no stipulation as to whether this classical information should be public, so that any third party might have access to it, or private, so that any third party does not have access.

This chapter introduces the private classical capacity theorem, which gives the maximum rate at which Alice can communicate classical information privately to Bob without anyone else in the universe knowing what she sent to him. The information processing task corresponding to this theorem was one of the earliest studied in quantum information theory, with the Bennett-Brassard-84 quantum key distribution protocol being the first proposed protocol for exploiting quantum mechanics to establish a shared secret key between two parties. The private classical capacity theorem is important for quantum key distribution because it establishes the maximum rate at which two parties can establish a shared secret key.

Another equally important, but less obvious utility of private classical communication is in establishing a protocol for quantum communication at the coherent information rate. Section 21.2 demonstrated a somewhat roundabout way of arriving at the conclusion that it is possible to communicate quantum information reliably at the coherent information rate—recall that we "coherified" the entanglement-assisted classical capacity theorem and then exploited the coherent communication identity and catalytic use of entanglement. Establishing achievability of the coherent information rate via private classical coding is another way of arriving at the same result, with the added benefit that the resulting protocol does not require the catalytic use of entanglement.

The intuition for quantum communication via privacy arises from the no-cloning theorem. Suppose that Alice is able to communicate private classical messages to Bob, so that the channel's environment (Eve) is not able to distinguish which message Alice is transmitting

to Bob. That is, Eve's state is completely independent of Alice's message if the transmitted message is private. Then we might expect it to be possible to make a coherent version of this private classical code by exploiting superpositions of the private classical codewords. Since Eve's states are independent of the quantum message that Alice is sending through the channel, she is not able to "steal" any of the coherence in Alice's superposed states. Given that the overall evolution of the channel to Bob and Eve is unitary and the fact that Eve does not receive any quantum information with this scheme, we should expect that the quantum information appears at the receiving end of the channel so that Bob can decode it. Were Eve able to obtain any information about the private classical messages, then Bob would not be able to decode all of the quantum information when they construct a coherent version of this private classical code. Otherwise, they would violate the no-cloning theorem. We discuss this important application of private classical communication in the next chapter.

This chapter follows a similar structure as previous chapters. We first detail the information processing task for private classical communication. Section 22.2 then states the private classical capacity theorem, with the following two sections proving the achievability part and the converse part. We end with a general discussion of the private classical capacity and a brief overview of the secret-key-assisted private classical capacity.

22.1 The Information Processing Task

We begin by describing the information processing task for private classical communication (we define an $(n, P - \delta, \epsilon)$ private classical code). Alice selects a message m uniformly at random from a set $\mathcal{M}$ of messages, and she also selects an index k uniformly at random from a set $\mathcal{K}$ to assist in randomizing Eve's information (thus, the set $\mathcal{K}$ is a *privacy amplification set*).[1] Let M and K denote the random variables corresponding to Alice's choice of m and k, respectively. Alice prepares some state $\rho_{m,k}^{A'^n}$ as input to many uses of the quantum channel $\mathcal{N}^{A'\to B}$. Randomizing over the K variable leads to the following state at the channel input:

$$\rho_m^{A'^n} \equiv \frac{1}{|\mathcal{K}|} \sum_{k\in\mathcal{K}} \rho_{m,k}^{A'^n}. \tag{22.1}$$

Alice transmits $\rho_{m,k}^{A'^n}$ over many independent uses of the channel, producing the following state at Bob's receiving end:

$$\mathcal{N}^{A'^n\to B^n}(\rho_{m,k}^{A'^n}), \tag{22.2}$$

where $\mathcal{N}^{A'^n\to B^n} \equiv (\mathcal{N}^{A'\to B})^{\otimes n}$.

Bob employs a decoding POVM $\{\Lambda_{m,k}\}$ in order to detect Alice's transmitted message m and the randomization variable k. The probability of error for a particular pair (m, k) is

[1] By convention, we usually assume that Alice picks her message M uniformly at random, though this is not strictly necessary. In the case of the randomizing variable K, it is required for Alice to select it uniformly at random in order to randomize Eve's knowledge of the Alice's message M. Otherwise, Alice's message M will not be indistinguishable to Eve.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

as follows:

$$p_e(m,k) = \text{Tr}\Big\{(I - \Lambda_{m,k})\mathcal{N}^{A'^n \to B^n}(\rho_{m,k}^{A'^n})\Big\}, \tag{22.3}$$

so that the maximal probability of error is

$$p_e^* \equiv \max_{m\in\mathcal{M},k\in\mathcal{K}} p_e(m,k). \tag{22.4}$$

The rate P of this code is

$$P \equiv \frac{1}{n}\log_2|\mathcal{M}| + \delta, \tag{22.5}$$

where δ is some arbitrarily small positive number.

So far, the above specification of a private classical code is nearly identical to that for the transmission of classical information outlined in Section 19.2. What distinguishes a private classical code from a public one is the following extra condition for privacy. Let $U_{\mathcal{N}}^{A'\to BE}$ be an isometric extension of the channel $\mathcal{N}^{A'\to B}$, so that the complementary channel $\widehat{\mathcal{N}}^{A'\to E}$ to the environment Eve is as follows:

$$\widehat{\mathcal{N}}^{A'\to E}(\sigma) \equiv \text{Tr}_B\{U_{\mathcal{N}}\sigma U_{\mathcal{N}}^\dagger\}. \tag{22.6}$$

If Alice transmits a message m, while selecting the variable k uniformly at random in order to randomize Eve's knowledge of the message m, then the expected state for Eve is as follows:

$$\omega_m^{E^n} \equiv \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} \widehat{\mathcal{N}}^{A'^n\to E^n}(\rho_{k,m}^{A'^n}). \tag{22.7}$$

Our condition for ϵ-privacy is that Eve's state is always close to a constant state, regardless of which message m Alice transmits through the channel:

$$\forall m \in \mathcal{M} : \left\|\omega_m^{E^n} - \omega^{E^n}\right\|_1 \leq \epsilon. \tag{22.8}$$

This definition is the strongest definition of privacy because it implies that Eve cannot learn anything about the message m that Alice transmits through the channel. Let σ^{E^n} denote Eve's state averaged over all possible messages:

$$\sigma^{E^n} \equiv \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \omega_m^{E^n}. \tag{22.9}$$

It follows that

$$\left\|\sigma^{E^n} - \omega^{E^n}\right\|_1 \leq \epsilon, \tag{22.10}$$

by applying convexity of the trace distance to (22.8). The criterion in (22.8) implies that Eve's Holevo information with M is arbitrarily small:

$$\begin{aligned} I(M;E^n) &= H(E^n) - H(E^n|M) && (22.11)\\ &= H\big(\sigma^{E^n}\big) - \frac{1}{|\mathcal{M}|}\sum_m H\big(\omega_m^{E^n}\big) && (22.12)\\ &\leq H\big(\omega^{E^n}\big) - \frac{1}{|\mathcal{M}|}\sum_m H\big(\omega^{E^n}\big) + 4\epsilon n\log d_E + 4H_2(\epsilon) && (22.13)\\ &= 4\epsilon n\log d_E + 4H_2(\epsilon). && (22.14)\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

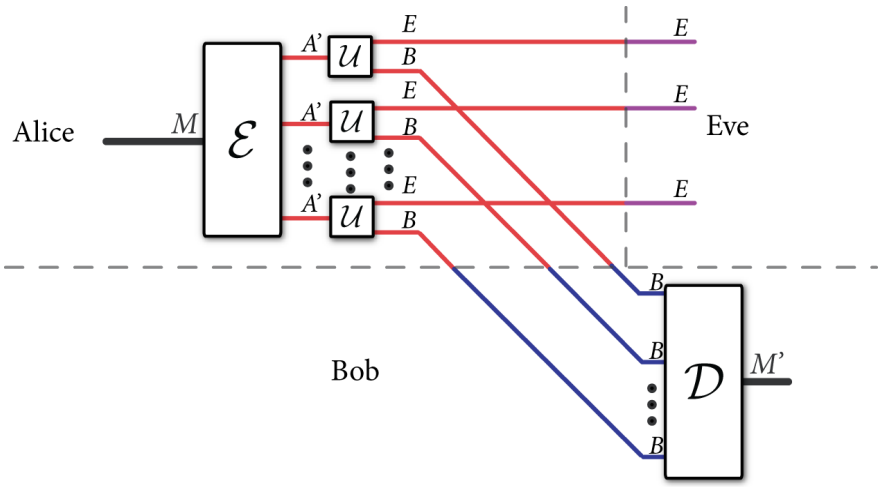

Figure 22.1: The information processing task for private classical communication. Alice encodes some private message m into a quantum codeword $\rho_m^{A'^n}$ and transmits it over many uses of a quantum channel. The goal of such a protocol is for Bob to be able to reliably distinguish the message, while the channel's environment Eve should not be able to learn anything about it.

The inequality follows from applying Fannes' inequality (Theorem 11.9.5) to both entropies. Thus, if ϵ is exponentially small in n (which will be the case for our codes), then it is possible to make Eve's information about the message become arbitrarily small in the asymptotic limit. Figure 22.1 depicts the information processing task for private classical communication.

In summary, we say that a rate P of private classical communication is achievable if there exists an $(n, P - \delta, \epsilon)$ private classical code for all $\epsilon, \delta > 0$ and sufficiently large n, where ϵ characterizes both the reliability and the privacy of the code.

22.2 The Private Classical Capacity Theorem

We now state the main theorem of this chapter, the private classical capacity theorem.

Theorem 22.2.1 (Devetak-Cai-Winter-Yeung)**.** *The private classical capacity of a quantum channel $\mathcal{N}^{A'\to B}$ is the supremum over all achievable rates for private classical communication, and one characterization of it is the regularization of the private information of the channel:*

$$\sup\{P \mid P \text{ is achievable}\} = P_{reg}(\mathcal{N}), \tag{22.15}$$

where

$$P_{reg}(\mathcal{N}) \equiv \lim_{k\to\infty} \frac{1}{k} P\left(\mathcal{N}^{\otimes k}\right). \tag{22.16}$$

The private information $P(\mathcal{N})$ is defined as

$$P(\mathcal{N}) \equiv \max_{\rho} [I(X;B)_\sigma - I(X;E)_\sigma], \tag{22.17}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $\rho^{XA'}$ is a classical-quantum state of the following form:

$$\rho^{XA'} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^{A'}, \tag{22.18}$$

and $\sigma^{XBE} \equiv U_{\mathcal{N}}^{A'\to BE}(\rho^{XA'})$, with $U_{\mathcal{N}}^{A'\to BE}$ an isometric extension of the channel $\mathcal{N}^{A'\to B}$.

We first prove the achievability part of the coding theorem and follow with the converse proof. Recall that the private information is additive whenever the channel is degradable (Theorem 12.6.3). Thus, for this class of channels, the regularization in (22.16) is not necessary and the private information of the channel is equal to the private classical capacity (in fact, the results from Theorem 12.6.2 and the next chapter demonstrate that the private information of a degradable channel is also equal to its quantum capacity). Unfortunately, it is known in the general case that the regularization of the private information is necessary in order to characterize the private capacity because there is an example of a channel for which the private information is superadditive.

22.3 The Direct Coding Theorem

This section gives the proof that the private information in (22.17) is an achievable rate for private classical communication over a quantum channel $\mathcal{N}^{A'\to B}$. We first give the intuition behind the protocol. Alice's goal is to build a doubly-indexed codebook $\{x^n(m,k)\}_{m\in\mathcal{M},k\in\mathcal{K}}$ that satisfies two properties:

1. Bob should be able to detect the message m and the "junk" variable k with high probability. From the classical coding theorem of Chapter 19, our intuition is that he should be able to do so as long as $|\mathcal{M}||\mathcal{K}| \approx 2^{nI(X;B)}$.

2. Randomizing over the "junk" variable k should approximately cover the typical subspace of Eve's system, so that every state of Eve depending on the message m looks like a constant, independent of the message m Alice sends (we would like the code to satisfy (22.8)). Our intuition from the Covering Lemma (Chapter 16) is that the size of the "junk" variable set $\mathcal{K}$ needs to be at least $|\mathcal{K}| \approx 2^{nI(X;E)}$ in order for Alice to approximately cover Eve's typical subspace.

Our method for generating a code is again of course random because we can invoke the typicality properties that hold in the asymptotic limit of many channel uses. Thus, if Alice chooses a code that satisfies the above criteria, she can send approximately $|\mathcal{M}| \approx 2^{n[I(X;B)-I(X;E)]}$ distinguishable signals to Bob such that they are indistinguishable to Eve. We devote the remainder of this section to proving that the above intuition is correct. Figure 22.2 displays the anatomy of a private classical code.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

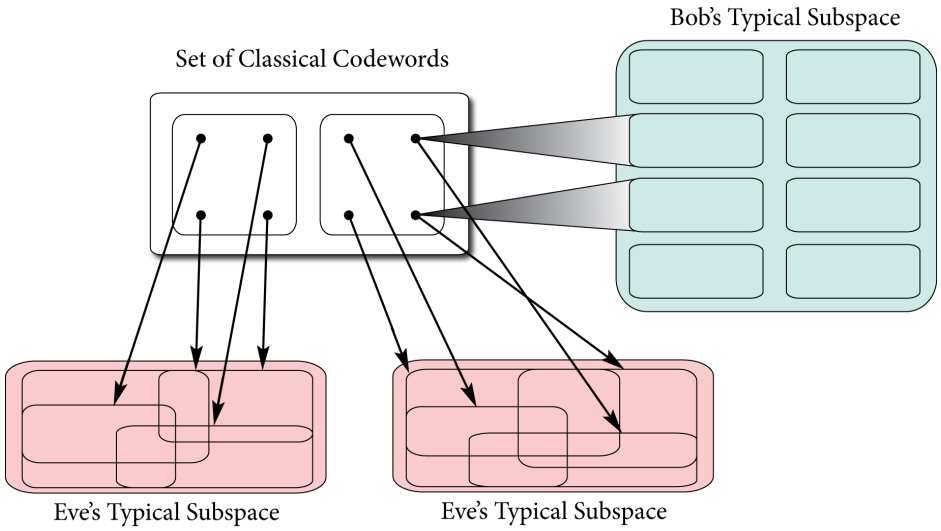

Figure 22.2: The anatomy of a code for private classical communication. In this illustrative example, Alice has eight codewords, with each depicted as a $\bullet$ and indexed by $m \in \{1, 2\}$ and $k \in \{1, 2, 3, 4\}$. Thus, she is interested in sending one of two messages and has the "junk" variable k available for randomizing Eve's state. Each classical codeword $x^n(m, k)$ maps to a distinguishable subspace on Bob's typical subspace (we show two of the mappings in the figure, while displaying eight distinguishable subspaces). From the Packing Lemma, our intuition is that Alice can reliably send about $2^{nI(X;B)}$ distinguishable signals. The codewords $\{x^n(1, k)\}_{k\in\{1,2,3,4\}}$ and $\{x^n(2, k)\}_{k\in\{1,2,3,4\}}$ are each grouped in a box to indicate that they form a privacy amplification set. When randomizing k, the codewords $\{x^n(1, k)\}_{k\in\{1,2,3,4\}}$ uniformly cover Eve's typical subspace (and so does the set $\{x^n(2, k)\}_{k\in\{1,2,3,4\}}$), so that it becomes nearly impossible in the asymptotic limit for Eve to distinguish whether Alice is sending a codeword in $\{x^n(1, k)\}_{k\in\{1,2,3,4\}}$ or $\{x^n(2, k)\}_{k\in\{1,2,3,4\}}$. In this way, Eve cannot determine which message Alice is transmitting. The minimum size for each privacy amplification set in the asymptotic limit is $\approx 2^{nI(X;E)}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Party	Quantity	Typical Set/Subspace	Projector
Alice	X	$T_\delta^{X^n}$	N/A
Bob	ρ^{B^n}	$T_\delta^{B^n}$	$\Pi_\delta^{B^n}$
Bob conditioned on x^n	$\rho_{x^n}^{B^n}$	$T_\delta^{B^n\|x^n}$	$\Pi_\delta^{B^n\|x^n}$
Eve	ρ^{E^n}	$T_\delta^{E^n}$	$\Pi_\delta^{E^n}$
Eve conditioned on x^n	$\rho_{x^n}^{E^n}$	$T_\delta^{E^n\|x^n}$	$\Pi_\delta^{E^n\|x^n}$

Table 22.1: The above table lists several mathematical quantities involved in the construction of a random private code. The first column lists the party to whom the quantities belong. The second column lists the random classical or quantum states. The third column gives the appropriate typical set or subspace. The final column lists the apprpriate projector onto the typical subspace for the quantum states.

22.3.1 Dimensionality Arguments

Before giving the proof of achievability, we confirm the above intuition with some dimensionality arguments and show how to satisfy the conditions of both the Packing and Covering Lemmas. Suppose that Alice has some ensemble $\{p_X(x), \rho_x^{A'}\}$ from which she can generate random codes. Let $U_{\mathcal{N}}^{A'\to BE}$ denote the isometric extension of the channel $\mathcal{N}^{A'\to B}$, and let ρ_x^{BE} denote the joint state of Bob and Eve after Alice inputs $\rho_x^{A'}$:

$$\rho_x^{BE} \equiv U_{\mathcal{N}} \rho_x^{A'} U_{\mathcal{N}}^\dagger. \tag{22.19}$$

The local respective density operators for Bob and Eve given a message x are as follows:

$$\rho_x^B \equiv \text{Tr}_E\{\rho_x^{BE}\}, \qquad \rho_x^E \equiv \text{Tr}_B\{\rho_x^{BE}\}. \tag{22.20}$$

The expected respective density operators for Bob and Eve are as follows:

$$\rho^B = \sum_x p_X(x)\rho_x^B, \qquad \rho^E = \sum_x p_X(x)\rho_x^E. \tag{22.21}$$

Given a particular input sequence x^n, we define the n^{th} extensions of the above states as follows:

$$\rho_{x^n}^{B^n} \equiv \text{Tr}_{E^n}\{\rho_{x^n}^{B^nE^n}\}, \tag{22.22}$$

$$\rho_{x^n}^{E^n} \equiv \text{Tr}_{B^n}\{\rho_{x^n}^{B^nE^n}\}, \tag{22.23}$$

$$\rho^{B^n} = \sum_{x^n\in\mathcal{X}^n} p_{X^n}(x^n)\rho_{x^n}^{B^n}, \tag{22.24}$$

$$\rho^{E^n} = \sum_{x^n\in\mathcal{X}^n} p_{X^n}(x^n)\rho_{x^n}^{E^n}. \tag{22.25}$$

The following four conditions corresponding to the Packing Lemma hold for Bob's states $\{\rho_{x^n}^{B^n}\}$, Bob's average density operator ρ^{B^n}, Bob's typical subspace $T_\delta^{B^n}$, and Bob's condi-

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

tionally typical subspace $T_\delta^{B^n|x^n}$:

$$\text{Tr}\{\rho_{x^n}^{B^n}\Pi_\delta^{B^n}\} \geq 1-\epsilon, \tag{22.26}$$

$$\text{Tr}\left\{\rho_{x^n}^{B^n}\Pi_\delta^{B^n|x^n}\right\} \geq 1-\epsilon, \tag{22.27}$$

$$\text{Tr}\left\{\Pi_\delta^{B^n|x^n}\right\} \leq 2^{n(H(B|X)+c\delta)}, \tag{22.28}$$

$$\Pi_\delta^{B^n}\rho^{B^n}\Pi_\delta^{B^n} \leq 2^{-n(H(B)-c\delta)}\Pi_\delta^{B^n}, \tag{22.29}$$

where c is some positive constant (see Properties 14.2.7, 14.1.3, 14.1.2, and 14.1.1).

The following four conditions corresponding to the Covering Lemma hold for Eve's states $\{\rho_{x^n}^{E^n}\}$, Eve's typical subspace $T_\delta^{E^n}$, and Eve's conditionally typical subspace $T_\delta^{E^n|x^n}$:

$$\text{Tr}\{\rho_{x^n}^{E^n}\Pi_\delta^{E^n}\} \geq 1-\epsilon, \tag{22.30}$$

$$\text{Tr}\left\{\rho_{x^n}^{E^n}\Pi_\delta^{E^n|x^n}\right\} \geq 1-\epsilon, \tag{22.31}$$

$$\text{Tr}\{\Pi_\delta^{E^n}\} \leq 2^{n(H(E)+c\delta)}, \tag{22.32}$$

$$\Pi_\delta^{E^n|x^n}\rho_{x^n}^{E^n}\Pi_\delta^{E^n|x^n} \leq 2^{-n(H(E|X)-c\delta)}\Pi_\delta^{E^n|x^n}. \tag{22.33}$$

The above properties suggest that we can use the methods of both the Packing Lemma and the Covering Lemma for constructing a private code. Consider two sets $\mathcal{M}$ and $\mathcal{K}$ with the following respective sizes:

$$|\mathcal{M}| = 2^{n[I(X;B)-I(X;E)-6c\delta]}, \tag{22.34}$$

$$|\mathcal{K}| = 2^{n[I(X;E)+3c\delta]}, \tag{22.35}$$

so that the product set $\mathcal{M}\times\mathcal{K}$ indexed by the ordered pairs (m,k) is of size

$$|\mathcal{M}\times\mathcal{K}| = |\mathcal{M}||\mathcal{K}| = 2^{n[I(X;B)-3c\delta]}. \tag{22.36}$$

The sizes of these sets suggest that we can use the product set $\mathcal{M}\times\mathcal{K}$ for sending classical information, but we can use $|\mathcal{M}|$ "privacy amplification" sets each of size $|\mathcal{K}|$ for reducing Eve's knowledge of the message m (see Figure 22.2).

22.3.2 Random Code Construction

We now argue for the existence of a good private classical code with rate $P = I(X;B) - I(X;E)$ if Alice selects it randomly according to the ensemble

$$\{p'_{X'^n}(x^n), \rho_{x^n}^{A'}\}, \tag{22.37}$$

where $p'_{X'^n}(x^n)$ is the pruned distribution (see Section 19.3.1—recall that this distribution is close to the IID distribution). Let us choose $|\mathcal{M}||\mathcal{K}|$ random variables $X^n(m,k)$ according to the distribution $p'_{X'^n}(x^n)$ where the realizations of the random variables $X^n(m,k)$ take

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

values in $\mathcal{X}^n$. After selecting these codewords randomly, the code $\mathcal{C} = \{x^n(m,k)\}_{m\in\mathcal{M},k\in\mathcal{K}}$ is then a fixed set of codewords $x^n(m,k)$ depending on the message m and the randomization variable k.

We first consider how well Bob can distinguish the pair (m,k) and argue that the random code is a good code in the sense that the expectation of the average error probability over all codes is low. The Packing Lemma is the basis of our argument. By applying the Packing Lemma (Lemma 15.3.1) to (22.26-22.29), there exists a POVM $(\Lambda_{m,k})_{(m,k)\in\mathcal{M}\times\mathcal{K}}$ corresponding to the random choice of code that reliably distinguishes the states $\{\rho^{B^n}_{X^n(m,k)}\}_{m\in\mathcal{M},k\in\mathcal{K}}$ in the following sense:

$$\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\} = 1 - \mathbb{E}_{\mathcal{C}}\left\{\frac{1}{|\mathcal{M}||\mathcal{K}|}\sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}}\text{Tr}\{\rho^{B^n}_{X^n(m,k)}\Lambda_{m,k}\}\right\} \tag{22.38}$$

$$\leq 2\left(\epsilon + \sqrt{8\epsilon}\right) + 4\left(\frac{2^{n(H(B)-c\delta)}}{2^{n(H(B|X)+c\delta)}|\mathcal{M}\times\mathcal{K}|}\right)^{-1} \tag{22.39}$$

$$= 2\left(\epsilon + \sqrt{8\epsilon}\right) + 4\left(\frac{2^{n(H(B)-c\delta)}}{2^{n(H(B|X)+c\delta)}2^{n[I(X;B)-3c\delta]}}\right)^{-1} \tag{22.40}$$

$$= 2\left(\epsilon + \sqrt{8\epsilon}\right) + 4\cdot 2^{-nc\delta} \equiv \epsilon'. \tag{22.41}$$

where the first equality follows by definition, the first inequality follows by application of the Packing Lemma to the conditions in (22.26-22.29), the second equality follows by substitution of (22.36), and the last equality follows by a straightforward calculation. We can make ϵ' arbitrarily small by choosing n large enough.

Let us now consider the corresponding density operators $\rho^{E^n}_{X^n(m,k)}$ for Eve. Consider dividing the random code $\mathcal{C}$ into $|\mathcal{M}|$ privacy amplification sets each of size $|\mathcal{K}|$. Each privacy amplification set $\mathcal{C}_m \equiv \{\rho^{E^n}_{X^n(m,k)}\}_{k\in\mathcal{K}}$ of density operators forms a good covering code according to the Covering Lemma (Lemma 16.2.1). The fake density operator of each privacy amplification set $\mathcal{C}_m$ is as follows:

$$\hat{\rho}^{E^n}_m \equiv \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\rho^{E^n}_{X^n(m,k)}, \tag{22.42}$$

because Alice chooses the randomizing variable k uniformly at random. The obfuscation error $o_e(\mathcal{C}_m)$ of each privacy amplification set $\mathcal{C}_m$ is as follows:

$$o_e(\mathcal{C}_m) \equiv \left\|\hat{\rho}^{E^n}_m - \rho^{E^n}\right\|_1, \tag{22.43}$$

where ρ^{E^n} is defined in (22.25). The Covering Lemma (Lemma 16.2.1) states that the obfuscation error for each random privacy amplification set $\mathcal{C}_m$ has a high probability of

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

being small if n is sufficiently large and $|\mathcal{K}|$ is chosen as in (22.35):

$$\begin{aligned}
&\Pr\{o_e(\mathcal{C}_m) \leq \epsilon + 4\sqrt{\epsilon} + 24\sqrt[4]{\epsilon}\} \\
&\geq 1 - 2d_E^n \exp\left\{ \frac{-\epsilon^3}{4\ln 2} \frac{|\mathcal{K}| 2^{n(H(E|X)-c\delta)}}{2^{n(H(E)+c\delta)}} \right\} && (22.44) \\
&= 1 - 2d_E^n \exp\left\{ \frac{-\epsilon^3}{4\ln 2} \frac{2^{n[I(X;E)+3c\delta]} 2^{n(H(E|X)-c\delta)}}{2^{n(H(E)+c\delta)}} \right\} && (22.45) \\
&= 1 - 2d_E^n \exp\left\{ \frac{-\epsilon^3}{4\ln 2} 2^{nc\delta} \right\}. && (22.46)
\end{aligned}$$

In particular, let us choose n large enough so that the following bound holds:

$$\Pr\{o_e(\mathcal{C}_m) \leq \epsilon + 4\sqrt{\epsilon} + 24\sqrt[4]{\epsilon}\} \geq 1 - \frac{\epsilon}{|\mathcal{M}|}. \tag{22.47}$$

That we can do so follows from the important fact that $\exp\{-\epsilon^3 2^{nc\delta}/(4\ln 2)\}$ is doubly exponentially decreasing in n. (We also see here why it is absolutely necessary to have the "wiggle room" given by an arbitrarily small, yet strictly positive δ.)

This random construction already has some of the desirable features that we are looking for in a private code just by choosing n to be sufficiently large. The expectation of Bob's average error probability for detecting the pair m, k is small, and the obfuscation error of each privacy amplification set has a high probability of being small. Our hope is that there exists some code for which Bob can retrieve the message m with the guarantee that Eve's knowledge is independent of this message m. We argue in the next two sections that such a good private code exists.

22.3.3 Derandomization

We now apply a derandomization argument similar to the one that is needed in the proof of the HSW coding theorem. The argument in this case is more subtle because we would like to find a code that has good classical communication with the guarantee that it also has good privacy. We need to determine the probability over all codes that there exists a good private code. If this probability is non-zero, then we are sure that a good private code exists.

As we have said at the beginning of this section, a good private code has two qualities: the code is ϵ-good for classical communication and it is ϵ-private as well. Let E_0 denote the event that the random code $\mathcal{C}$ is ϵ-good for classical communication:

$$E_0 = \{\bar{p}_e(\mathcal{C}) \leq \epsilon\}, \tag{22.48}$$

where we restrict the performance criterion to the average probability of error for now. Let E_m denote the event that the m^{th} message in the random code is ϵ-private:

$$E_m = \{o_e(\mathcal{C}_m) \leq \epsilon\}. \tag{22.49}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We would like all of the above events to be true, or, equivalently, we would like the intersection of the above events to occur:

$$E_{\text{priv}} \equiv E_0 \cap \bigcap_{m \in \mathcal{M}} E_m. \tag{22.50}$$

If there is a positive probability over all codes that the above event is true, then there exists a particular code that satisfies the above conditions. Let us instead consider the complement of the above event (the event that a good private code does not exist):

$$E_{\text{priv}}^c = E_0^c \cup \bigcup_{m \in \mathcal{M}} E_m^c. \tag{22.51}$$

We can then exploit the union bound from probability theory to bound the probability of the complementary event E_{priv}^c as follows:

$$\Pr\left\{ E_0^c \cup \bigcup_{m \in \mathcal{M}} E_m^c \right\} \leq \Pr\{E_0^c\} + \sum_{m \in \mathcal{M}} \Pr\{E_m^c\}. \tag{22.52}$$

So if we can make the probability of the event E_{priv}^c small, then the probability of the event E_{priv} that there exists a good private code is high.

Let us first bound the probability of the event E_0^c. Markov's inequality states that the following holds for a non-negative random variable Y:

$$\Pr\{Y \geq \alpha\} \leq \frac{\mathbb{E}\{Y\}}{\alpha}. \tag{22.53}$$

We can apply Markov's inequality because the random average error probability $\bar{p}_e(\mathcal{C})$ is always non-negative:

$$\Pr\{E_0^c\} = \Pr\left\{ \bar{p}_e(\mathcal{C}) \geq (\epsilon')^{3/4} \right\} \leq \frac{\mathbb{E}_{\mathcal{C}}\{\bar{p}_e(\mathcal{C})\}}{(\epsilon')^{3/4}} \leq \frac{\epsilon'}{(\epsilon')^{3/4}} = \sqrt[4]{\epsilon'} \tag{22.54}$$

So we now have a good bound on the probability of the complementary event E_0^c.

Let us now bound the probability of the events E_m^c. The bounds in the previous section already give us what we need:

$$\Pr\{E_m^c\} = \Pr\left\{ o_e(\mathcal{C}_m) > \epsilon + 4\sqrt{\epsilon} + 24\sqrt[4]{\epsilon} \right\} \tag{22.55}$$

$$< \frac{\epsilon}{|\mathcal{M}|}. \tag{22.56}$$

implying that

$$\sum_{m \in \mathcal{M}} \Pr\{E_m^c\} < |\mathcal{M}| \frac{\epsilon}{|\mathcal{M}|} = \epsilon. \tag{22.57}$$

So it now follows that the probability of the complementary event is small:

$$\Pr\left\{ E_{\text{priv}}^c \right\} \leq \sqrt[4]{\epsilon'} + \epsilon, \tag{22.58}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and there is a high probability that there is a good code:

$$\Pr\{E_{\text{priv}}\} \geq 1 - \left(\sqrt[4]{\epsilon'} + \epsilon\right). \tag{22.59}$$

Thus, there exists a particular code $\mathcal{C}$ such that its average probability of error is small for decoding the classical information:

$$\bar{p}_e(\mathcal{C}) \leq (\epsilon')^{3/4}, \tag{22.60}$$

and the obfuscation error of each privacy amplification set is small:

$$\forall m : o_e(\mathcal{C}_m) \leq \epsilon + 4\sqrt{\epsilon} + 24\sqrt[4]{\epsilon}. \tag{22.61}$$

The derandomized code $\mathcal{C}$ is as follows:

$$\mathcal{C} \equiv \{x^n(m,k)\}_{m\in\mathcal{M},k\in\mathcal{K}}, \tag{22.62}$$

so that each codeword $x^n(m,k)$ is a deterministic variable. Each privacy amplification set for the derandomized code is as follows:

$$\mathcal{C}_m \equiv \{x^n(m,k)\}_{k\in\mathcal{K}}. \tag{22.63}$$

The result in (22.59) is perhaps astonishing in hindsight. By choosing a private code in a random way and choosing the block length n of the private code to be sufficiently large, the overwhelming majority of codes constructed in this fashion are good private codes!

22.3.4 Expurgation

We would like to strengthen the above result even more, so that the code has a low maximal probability of error, not just a low average error probability. We expurgate codewords from the code as before, but we have to be careful with the expurgation argument because we need to make sure that the code still has good privacy after expurgation.

We can apply Markov's inequality for the expurgation in a way similar as in Exercise 2.2.1. It is possible to apply Markov's inequality to the bound on the average error probability in (22.54) to show that at most a fraction $\sqrt{\epsilon'}$ of the codewords have error probability greater than $\sqrt[4]{\epsilon'}$. We could merely expurgate the worst $\sqrt{\epsilon'}$ codewords from the private code. But expurgating in this fashion does not guarantee that each privacy amplification set has the same number of codewords. Therefore, we expurgate the worst fraction $\sqrt{\epsilon'}$ of the codewords in each privacy amplification set. We then expurgate the worst fraction $\sqrt{\epsilon'}$ of the privacy amplification sets. The expurgated sets $\mathcal{M}'$ and $\mathcal{K}'$ both become a fraction $1-\sqrt{\epsilon'}$ of their original size. We denote the expurgated code as follows:

$$\mathcal{C}' \equiv \{x^n(m,k)\}_{m\in\mathcal{M}',k\in\mathcal{K}'}, \tag{22.64}$$

and the expurgated code has the following privacy amplification sets:

$$\mathcal{C}'_m \equiv \{x^n(m,k)\}_{k\in\mathcal{K}'}. \tag{22.65}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The expurgation has a negligible impact on the rate of the private code when n is large.

Does each privacy amplification set still have good privacy properties after performing the above expurgation? The fake density operator for each expurgated privacy amplification set is as follows:

$$\hat{\rho}_m^{E^n\prime} \equiv \frac{1}{|\mathcal{C}'_m|} \sum_{k \in \mathcal{K}'} \rho_{x^n(m,k)}^{E^n}. \tag{22.66}$$

It is possible to show that the trace distance between the fake density operators in the derandomized code are $2\sqrt{\epsilon'}$-close in trace distance to the fake density operators in the expurgated code:

$$\forall m \in \mathcal{M}' \qquad \left\| \hat{\rho}_m^{E^n\prime} - \hat{\rho}_m^{E^n} \right\|_1 \leq 2\sqrt{\epsilon'}, \tag{22.67}$$

because these operators only lose a small fraction of their mass after expurgation.

We now drop the primed notation for the expurgated code. It follows that the expurgated code $\mathcal{C}$ has good privacy:

$$\forall m \in \mathcal{M} \qquad \left\| \hat{\rho}_m^{E^n} - \rho^{E^n} \right\|_1 \leq \epsilon + 4\sqrt{\epsilon} + 24\sqrt[4]{\epsilon} + 2\sqrt{\epsilon'}, \tag{22.68}$$

and reliable communication:

$$\forall m \in \mathcal{M},\ k \in \mathcal{K} \qquad p_e(\mathcal{C}, m, k) \leq \sqrt[4]{\epsilon'}. \tag{22.69}$$

The first expression follows by application of the triangle inequality to (22.61) and (22.67).

We end the proof by summarizing the operation of the private code. Alice chooses the message m and the randomization variable k uniformly at random from the respective sets $\mathcal{M}$ and $\mathcal{K}$. She encodes these as $x^n(m,k)$ and inputs the quantum codeword $\rho_{x^n(m,k)}^{A'^n}$ to the channel. Bob receives the state $\rho_{x^n(m,k)}^{B^n}$ and performs a POVM $(\Lambda_{m,k})_{(m,k)\in\mathcal{M}\times\mathcal{K}}$ that determines the pair m and k correctly with probability $1 - \sqrt[4]{\epsilon'}$. The code guarantees that Eve has almost no knowledge about the message m. The private communication rate P of the private code is equal to the following expression:

$$P \equiv \frac{1}{n} \log_2 |\mathcal{M}| = I(X;B) - I(X;E) - 6c\delta. \tag{22.70}$$

This concludes the proof of the direct coding theorem.

We remark that the above proof applies even in the scenario where Eve does not get the full purification of the channel. That is, suppose that the channel has one input A' for Alice and two outputs B and E for Bob and Eve, respectively. Then the channel has an isometric extension to some environment F. In this scenario, the private information $I(X;B) - I(X;E)$ is still achievable for some classical-quantum state input such that the Holevo information difference is positive. Though, one could always give both outputs E and F to an eavesdropper (this is the setting that we proved in the above theorem). Giving the full purification of the channel to the environment ensures that the transmitted information is private from the "rest of the universe" (anyone other than the intended receiver), and it thus yields the highest standard of security in any protocol for private information transmission.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

22.4 The Converse Theorem

We now prove the converse part of the private classical capacity theorem, which demonstrates that the regularization of the private information is an upper bound on the private classical capacity. We suppose instead that Alice and Bob are trying to accomplish the task of secret key generation. As we have argued in other converse proofs (see Sections 19.3.2 and 20.5), the capacity for generating this static resource can only be larger than the capacity for private classical communication because Alice and Bob can always use a noiseless private channel to establish a shared secret key. In such a task, Alice first prepares a maximally correlated state $\overline{\Phi}^{MM'}$ and encodes the M' variable as a codeword of the form in (22.1). This encoding leads to a state of the following form, after Alice transmits her systems A'^n over many independent uses of the channel:

$$\omega^{MB^nE^n} \equiv \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} |m\rangle\langle m|^M \otimes U_{\mathcal{N}}^{A'^n\to B^nE^n}(\rho_m^{A'^n}). \tag{22.71}$$

Bob finally applies a decoding map $\mathcal{D}^{B^n\to M'}$ to recover his share of the secret key:

$$\omega^{MM'E^n} \equiv \mathcal{D}^{B^n\to M'}(\omega^{MB^nE^n}). \tag{22.72}$$

If the code is good for secret key generation, then the following condition should hold

$$\left\|\omega^{MM'E^n} - \overline{\Phi}^{MM'}\otimes\sigma^{E^n}\right\|_1 \leq \epsilon, \tag{22.73}$$

so that Eve's state σ^{E^n} is a constant state independent of the secret key $\overline{\Phi}^{MM'}$. In particular, the above condition implies that Eve's information about M is small:

$$I(M;E^n)_\omega \leq n\epsilon', \tag{22.74}$$

where we apply the Alicki-Fannes' inequality to (22.73) with $\epsilon' \equiv 6\epsilon\log d_E + 4H_2(\epsilon)/n$. The rate $P-\delta$ of secret key generation is equal to $\frac{1}{n}\log|\mathcal{M}|$. Consider the following chain of inequalities:

$$\begin{aligned}
n(P-\delta) &= I(M;M')_{\overline{\Phi}} && (22.75)\\
&\leq I(M;M')_\omega + n\epsilon' && (22.76)\\
&\leq I(M;B^n)_\omega + n\epsilon' && (22.77)\\
&\leq I(M;B^n)_\omega - I(M;E^n)_\omega + 2n\epsilon' && (22.78)\\
&\leq P\big(\mathcal{N}^{\otimes n}\big) + 2n\epsilon'. && (22.79)
\end{aligned}$$

The first equality follows because the mutual information of the common randomness state $\overline{\Phi}^{MM'}$ is equal to $n(P-\delta)$. The first inequality follows from applying the Alicki-Fannes' inequality to (22.73) with the above choice of ϵ'. The second inequality is quantum data processing. The third inequality follows from (22.74), and the final inequality follows because the classical-quantum state in (22.71) has a particular distribution and choice of states, and this choice always leads to a value of the private information that cannot be larger than the private information of the tensor product channel $\mathcal{N}^{\otimes n}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Exercise 22.4.1 Prove that free access to a forward public classical channel from Alice to Bob cannot improve the private classical capacity of a quantum channel.

22.5 Discussion of Private Classical Capacity

This last section discusses some important aspects of the private classical capacity. Two of these results have to do with the fact that Theorem 22.2.1 only provides a regularized characterization of the private classical capacity, and the last asks what rates of private classical communication are achievable if the sender and receiver share a secret key before communication begins. For full details, we refer the reader to the original papers in the quantum Shannon theory literature.

22.5.1 Superadditivity of the Private Information

Theorem 22.2.1 states that the private classical capacity of a quantum channel is equal to the regularized private information of the channel. As we have said before (at the beginning of Chapter 20), a regularized formula is not particularly useful from a practical perspective because it is impossible to perform the optimization task that it sets out, and it is not desirable from an information-theoretical perspective because such a regularization does not identify a formula as a unique measure of correlations.

In light of the unsatisfactory nature of a regularized formula, is it really necessary to have the regularization in Theorem 22.2.1 for arbitrary quantum channels? Interestingly, the answer is "yes" in the general case (though, we know it is not necessary if the channel is degradable). The reason is that there exists an example of a channel $\mathcal{N}$ for which the private information is strictly superadditive:

$$mP(\mathcal{N}) < P\big(\mathcal{N}^{\otimes m}\big), \tag{22.80}$$

for some positive integer m. Specifically, Smith *et al.* showed that the private information of a particular Pauli channel exhibits this superadditivity [231]. To do so, they calculated the private information $P(\mathcal{N})$ for such a channel. Next, they consider performing an m-qubit "repetition code" before transmitting qubits into the channel. A repetition code is a quantum code that performs the following encoding:

$$\alpha|0\rangle + \beta|1\rangle \rightarrow \alpha|0\rangle^{\otimes m} + \beta|1\rangle^{\otimes m}. \tag{22.81}$$

Evaluating the private information when sending a particular state through the repetition code and then through m instances of the channel leads to a higher value than $mP(\mathcal{N})$, implying the strict inequality in (22.80). Thus, additivity of the private information formula $P(\mathcal{N})$ cannot hold in the general case.

The implications of this result are that we really do not understand the best way of transmitting information privately over a quantum channel that is not degradable, and it is thus the subject of ongoing research.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

22.5.2 Superadditivity of Private Classical Capacity

The private information of a particular channel can be superadditive (as discussed in the previous section), and so the regularized private information is our best characterization of the capacity for this information processing task. In spite of this, we might hope that some eventual formula for the private classical capacity would be additive (some formula other than the private information $P(\mathcal{N})$). Interestingly, this is also not the case.

To clarify this point, suppose that $P^{?}(\mathcal{N})$ is some formula for the private classical capacity. If it were an additive formula, then it should be additive as a function of channels:

$$P^{?}(\mathcal{N} \otimes \mathcal{M}) = P^{?}(\mathcal{N}) + P^{?}(\mathcal{M}). \tag{22.82}$$

Li *et al.* have shown that this cannot be the case for any proposed private capacity formula, by making a clever argument with a construction of channels [183]. Specifically, they constructed a particular channel $\mathcal{N}$ which has a single-letter *classical* capacity. The fact that the channel's classical capacity is sharply upper bounded implies that its private classical capacity is as well. Let D be the upper bound so that $P^{?}(\mathcal{N}) \leq D$. Also, they considered a 50% erasure channel, one which gives the input state to Bob and an erasure symbol to Eve with probability 1/2 and gives the input state to Eve and an erasure symbol to Bob with probability 1/2. Such a channel has zero capacity for sending private classical information essentially because Eve is getting the same amount of information as Bob does on average. Thus, $P^{?}(\mathcal{M}) = 0$. In spite of this, Li *et al.* show that the tensor product channel $\mathcal{N} \otimes \mathcal{M}$ has a private classical capacity that exceeds D. We can then make the conclusion that these two channels allow for superadditivity of private classical capacity:

$$P^{?}(\mathcal{N} \otimes \mathcal{M}) > P^{?}(\mathcal{N}) + P^{?}(\mathcal{M}), \tag{22.83}$$

and that (22.82) cannot hold in the general case. More profoundly, their results demonstrate that the private classical capacity itself is non-additive, even if a characterization of it is found that is more desirable than that with the formula in Theorem 22.2.1. Thus, it will likely be difficult to obtain a desirable characterization of the private classical capacity for general quantum channels.

22.5.3 Secret-key Assisted Private Classical Communication

The direct coding part of Theorem 22.2.1 demonstrates how to send private classical information over a quantum channel $\mathcal{N}$ at the private information rate $P(\mathcal{N})$. A natural extension to consider is the scenario where Alice and Bob share a secret key before communication begins. A secret key shared between Alice and Bob and secure from Eve is a tripartite state of the following form:

$$\overline{\Phi}^{AB} \otimes \sigma^{E}, \tag{22.84}$$

where $\overline{\Phi}^{AB}$ is the maximally correlated state and σ^{E} is a state on Eve's system that is independent of the key shared between Alice and Bob. Like the entanglement-assisted capacity

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

theorem, we assume that they obtain this secret key from some third party, and the third party ensures that the key remains secure.

The resulting capacity theorem is known as the secret-key-assisted private classical capacity theorem, and it characterizes the trade-off between secret key consumption and private classical communication. The main idea for this setting is to show the existence of a protocol that transmits private classical information at a rate of $I(X;B)$ private bits per channel use while consuming secret key at a rate of $I(X;E)$ secret key bits per channel use, where the information quantities are with respect to the state in Theorem 22.2.1. The protocol for achieving these rates is almost identical to the one we gave in the proof of the direct coding theorem, though with one difference. Instead of sacrificing classical bits at a rate of $I(X;E)$ in order to randomize Eve's knowledge of the message (recall that our randomization variable had to be chosen uniformly at random from a set of size $\approx 2^{nI(X;E)}$), the sender exploits the secret key to do so. The converse proof shows that this strategy is optimal (with a multi-letter characterization). Thus, we have the following capacity theorem.

Theorem 22.5.1 (Secret-key-assisted capacity theorem)**.** *The secret-key-assisted private classical capacity region $C_{SKA}(\mathcal{N})$ of a quantum channel $\mathcal{N}$ is given by*

$$C_{SKA}(\mathcal{N}) = \overline{\bigcup_{k=1}^{\infty} \frac{1}{k} \widetilde{C}^{(1)}_{SKA}(\mathcal{N}^{\otimes k})}, \tag{22.85}$$

where the overbar indicates the closure of a set. $\widetilde{C}^{(1)}_{SKA}(\mathcal{N})$ is the set of all $P, S \geq 0$ such that

$$P \leq I(X;B)_{\sigma} - I(X;E)_{\sigma} + S, \tag{22.86}$$
$$P \leq I(X;B)_{\sigma}. \tag{22.87}$$

where P is the rate of private classical communication, S is the rate of secret key consumption, the state σ^{XBE} is of the following form

$$\sigma^{XBE} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes U_{\mathcal{N}}^{A' \to BE}\left(\rho_x^{A'}\right), \tag{22.88}$$

and $U_{\mathcal{N}}^{A' \to BE}$ is an isometric extension of the channel.

Showing that the above inequalities are achievable follows by time-sharing between the protocol from the direct coding part of Theorem 22.2.1 and the aforementioned protocol for secret-key-assisted private classical communication.

22.6 History and Further Reading

Bennett and Brassard devised the first protocol for sending private classical data over a quantum channel [22]. The protocol given there became known as quantum key distribution, which has now become a thriving field in its own right [212]. Devetak [68] and Cai,

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Winter, and Yeung [51] proved the characterization of the private classical capacity given in this chapter (both using the techniques in this chapter). Hsieh *et al.* [157] proved achievability of the secret-key-assisted protocol given in Section 22.5.3, and Ref. [248] proved the converse and stated the secret-key-assisted capacity theorem. Later work characterized the full trade-off between public classical communication, private classical communication, and secret key [158, 252]. Smith *et al.* showed that the private information can exhibit superadditivity [231], and Li *et al.* showed that the private classical capacity is generally nonadditive [183]. Smith later showed that the symmetric-side-channel-assisted private classical capacity is additive [230]. Datta and Hsieh recently demonstrated universal private codes for quantum channels [61].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 23

Quantum Communication

The quantum capacity theorem is one of the most important theorems in quantum Shannon theory. It is a fundamentally "quantum" theorem in that it demonstrates that a fundamentally quantum information quantity, the coherent information, is an achievable rate for quantum communication over a quantum channel. The fact that the coherent information does not have a strong analog in classical Shannon theory truly separates the quantum and classical theories of information.

The no-cloning theorem (Section 3.5.4) provides the intuition behind the quantum capacity theorem. The goal of any quantum communication protocol is for Alice to establish quantum correlations with the receiver Bob. We know well now that every quantum channel has an isometric extension, so that we can think of another receiver, the environment Eve, who is at a second output port of a larger unitary evolution. Were Eve able to learn anything about the quantum information that Alice is attempting to transmit to Bob, then Bob could not be retrieving this information—otherwise, they would violate the no-cloning theorem. Thus, Alice should figure out some subspace of the channel input where she can place her quantum information such that only Bob has access to it, while Eve does not. That the dimensionality of this subspace is exponential in the coherent information is perhaps then unsurprising in light of the above no-cloning reasoning. The coherent information is an entropy difference $H(B) - H(E)$—a measure of the amount of quantum correlations that Alice can establish with Bob less the amount that Eve can gain.[1]

We proved achievability of the coherent information for quantum data transmission in Corollary 21.2.1, but the roundabout path that we followed to prove achievability there perhaps does not give much insight into the structure of a quantum code that achieves the coherent information. Our approach in this chapter is different and should shed more light on this structure. Specifically, we show how to make coherent versions of the private classical codes from the previous chapter. By exploiting the privacy properties of these codes, we can form subspaces where Alice can store her quantum information such that Eve does not have access to it. Thus, this approach follows the above "no-cloning intuition" more closely.

[1]Recall from Exercise 11.6.6 that we can also write the coherent information as half the difference of Bob's mutual information with Alice less Eve's: $I(A\rangle B) = 1/2[I(A;B) - I(A;E)]$.

The best characterization that we have for the quantum capacity of a general quantum channel is the regularized coherent information. It turns out that the regularization is not necessary for the class of degradable channels, implying that we have a complete understanding of the quantum data transmission capabilities of these channels. However, if a channel is not degradable, there can be some startling consequences, and these results imply that we have an incomplete understanding of quantum data transmission in the general case. First, the coherent information can be strictly superadditive for the depolarizing channel. This means that the best strategy for achieving the quantum capacity is not necessarily the familiar one where we generate random quantum codes from a single instance of a channel. This result is also in marked constrast with the "classical" strategies that achieve the unassisted and entanglement-assisted classical capacities of the depolarizing channel. Second, perhaps the most surprising result in quantum Shannon theory is that it is possible to "superactivate" the quantum capacity. That is, suppose that two channels on their own have zero capacity for transmitting quantum information (for the phenomenon to occur, these channels are specific channels). Then it is possible for the joint channel (the tensor product of the individual channels) to have a non-zero quantum capacity, in spite of them being individually useless for quantum data transmission. This latter result implies that we are rather distant from having a complete quantum theory of information, in spite of the many successes reviewed in this book.

We structure this chapter as follows. We first overview the information processing task relevant for quantum communication. Next, we discuss the no-cloning intuition for quantum capacity in some more detail, presenting the specific example of a quantum erasure channel. Section 23.3 states the quantum capacity theorem, and the following two sections prove the direct coding and converse theorems corresponding to it. Section 23.6 computes the quantum capacity of two degradable channels: the quantum erasure channel and the amplitude damping channel. We then discuss superadditivity of coherent information and superactivation of quantum capacity in Section 23.7. Finally, we prove the existence of an entanglement distillation protocol, whose proof bears some similarities to the proof of the direct coding part of the quantum capacity theorem.

23.1 The Information Processing Task

We begin the technical development in this chapter by describing the information processing task for quantum communication (we define an $(n, Q - \delta, \epsilon)$ quantum communication code). First, there are several protocols that we can consider for quantum communication, but perhaps the strongest definition of quantum capacity corresponds to a task known as *entanglement transmission*. Suppose that Alice shares entanglement with a reference system to which she does not have access. Then their goal is to devise a quantum coding scheme such that Alice can transfer this entanglement to Bob. To this end, suppose that Alice and the reference share an arbitrary state $|\varphi\rangle^{RA_1}$. Alice then performs some encoder on system A_1 to prepare it for input to many instances of a quantum channel $\mathcal{N}^{A'\to B}$. The resulting

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

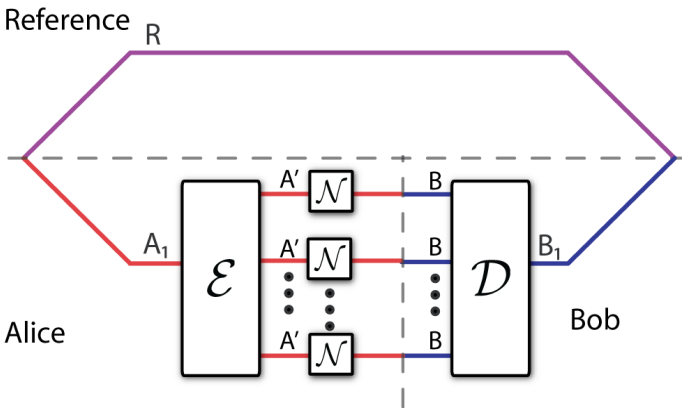

Figure 23.1: The information processing task for entanglement transmission. Alice is trying to preserve the entanglement with some inaccessible reference system by encoding her system and transmitting the encoded quantum data over many independent uses of a noisy quantum channel. Bob performs a decoding of the systems he receives, and the state at the end of the protocol is close to the original state shared between Alice and the reference if the protocol is any good for entanglement transmission.

state is as follows:

$$\mathcal{E}^{A_1 \to A'^n}(\varphi^{RA_1}). \tag{23.1}$$

Alice transmits the systems A'^n through many independent uses of the channel, resulting in the following state:

$$\mathcal{N}^{A'^n \to B^n}(\mathcal{E}^{A_1 \to A'^n}(\varphi^{RA_1})), \tag{23.2}$$

where $\mathcal{N}^{A'^n \to B^n} \equiv (\mathcal{N}^{A' \to B})^{\otimes n}$. After Bob receives the systems B^n from the channel outputs, he performs some decoding map $\mathcal{D}^{B^n \to B_1}$, where B_1 is some system of the same dimension as A_1. The final state after Bob decodes is as follows:

$$\omega^{RB_1} \equiv \mathcal{D}^{B^n \to B_1}(\mathcal{N}^{A'^n \to B^n}(\mathcal{E}^{A_1 \to A'^n}(\varphi^{RA_1}))). \tag{23.3}$$

Figure 23.1 depicts all of the above steps.

If the protocol is good for quantum communication, then the following condition should hold for all states $|\varphi\rangle^{RA_1}$:

$$\left\| \varphi^{RA_1} - \omega^{RB_1} \right\|_1 \leq \epsilon. \tag{23.4}$$

The rate Q of this scheme is equal to the number of qubits transmitted per channel use:

$$Q \equiv \frac{1}{n} \log d_{A_1} + \delta, \tag{23.5}$$

where d_{A_1} is the dimension of the A_1 register and δ is an arbitrarily small positive number. We say that a rate Q is achievable if there exists an $(n, Q - \delta, \epsilon)$ quantum communication code for all $\epsilon, \delta > 0$ and sufficiently large n.

The above notion of quantum communication encompasses other quantum information processing tasks such as mixed state transmission, pure state transmission, and entanglement

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

generation. Alice can transmit any mixed or pure state if she can preserve the entanglement with a reference system. Also, she can generate entanglement with Bob if she can preserve entanglement with a reference system—she just needs to create an entangled state locally and apply the above protocol to one system of the entangled state.

23.2 The No-Cloning Theorem and Quantum Communication

We first discuss quantum communication over a quantum erasure channel before stating and proving the quantum capacity theorem. Consider the quantum erasure channel that gives Alice's input state to Bob with probability $1-\epsilon$ and an erasure flag to Bob with probability ϵ:

$$\rho \rightarrow (1-\epsilon)\rho + \epsilon|e\rangle\langle e|, \tag{23.6}$$

where $\langle e|\rho|e\rangle = 0$ for all inputs ρ. Recall that the isometric extension of this channel is as follows (see Exercise 5.2.6):

$$|\psi\rangle^{RA} \rightarrow \sqrt{1-\epsilon}|\psi\rangle^{RB}|e\rangle^{E} + \sqrt{\epsilon}|\psi\rangle^{RE}|e\rangle^{B}, \tag{23.7}$$

so that the channel now has the other interpretation that Eve gets the state with probability ϵ while giving her the erasure flag with probability $1-\epsilon$.

Now suppose that the erasure parameter is set to $1/2$. In such a scenario, the channel to Eve is the *same* as the channel to Bob, namely, both have the channel $\rho \rightarrow 1/2(\rho + |e\rangle\langle e|)$. We can argue that the quantum capacity of such a channel should be zero, by invoking the no-cloning theorem. More specifically, suppose there is a scheme (an encoder and decoder as given in Figure 23.1) for Alice and Bob to communicate quantum information reliably at a non-zero rate over such a channel. If so, Eve could simply use the same decoder that Bob does, and she should also be able to obtain the quantum information that Alice is sending. But the ability for both Bob and Eve to decode the quantum information that Alice is transmitting violates the no-cloning theorem. Thus, the quantum capacity of such a channel should vanish.

Exercise 23.2.1 Prove that the quantum capacity of an amplitude damping channel vanishes if its damping parameter is equal to $1/2$.

The no-cloning theorem plays a more general role in the analysis of quantum communication over quantum channels. In the construction of a quantum code, we are trying to find a "no-cloning" subspace of the input Hilbert space that is protected from Eve. If Eve is able to obtain any of the quantum information in this subspace, then this information cannot be going to Bob by the same no-cloning argument featured in the previous paragraph. Thus, we might then suspect that the codes from the previous chapter for private classical communication might play a role for quantum communication because we constructed them in such a way that Eve would not be able to obtain any information about the private message

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

that Alice is transmitting to Eve. The main insight needed is to make a coherent version of these private classical codes, so that Alice and Bob conduct every step in superposition (much like we did in Chapter 21).

23.3 The Quantum Capacity Theorem

The main theorem of this chapter is the following quantum capacity theorem.

Theorem 23.3.1 (Quantum Capacity)**.** *The quantum capacity of a quantum channel $\mathcal{N}^{A'\to B}$ is the supremum over all achievable rates for quantum communication, and one characterization of it is the regularization of the coherent information of the channel:*

$$\sup\{Q \mid Q \text{ is achievable}\} = Q_{reg}(\mathcal{N}), \tag{23.8}$$

where

$$Q_{reg}(\mathcal{N}) \equiv \lim_{k\to\infty} \frac{1}{k} Q\big(\mathcal{N}^{\otimes k}\big). \tag{23.9}$$

The channel coherent information $Q(\mathcal{N})$ is defined as

$$Q(\mathcal{N}) \equiv \max_{\phi} I(A\rangle B)_{\sigma}, \tag{23.10}$$

where the optimization is over all pure, bipartite states $\phi^{AA'}$ and

$$\sigma^{AB} \equiv \mathcal{N}^{A'\to B}(\phi^{AA'}). \tag{23.11}$$

We prove this theorem in two parts: the direct coding theorem and the converse theorem. The proof of the direct coding theorem proceeds by exploiting the private classical codes from the previous chapter. The proof of the converse theorem is similar to approaches from previous chapters—we exploit the Alicki-Fannes' inequality and quantum data processing in order to obtain an upper bound on the quantum capacity. In general, the regularized coherent information is our best characterization of the quantum capacity, but the regularization is not necessary for the class of degradable channels. Since many channels of interest are degradable (including dephasing, amplitude damping, and erasure channels), we can calculate their quantum capacities.

23.4 The Direct Coding Theorem

The proof of the direct coding part of the quantum capacity theorem follows by taking advantage of the properties of the private classical codes constructed in the previous chapter (see Section 22.3). We briefly recall this construction. Suppose that a classical-quantum-quantum channel connects Alice to Bob and Eve. Specifically, if Alice inputs a classical letter x to the channel, then Bob receives a density operator ρ_x^B and Eve receives a density

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

operator ω_x^E. The direct coding part of Theorem 22.2.1 establishes the existence of a codebook $\{x^n(m,k)\}_{m\in\mathcal{M},k\in\mathcal{K}}$ selected from a distribution $p_X(x)$ and a corresponding decoding POVM $\{\Lambda_{m,k}^{B^n}\}$ such that Bob can detect Alice's message m and randomizing variable k with high probability:

$$\forall m \in \mathcal{M}, k \in \mathcal{K} : \text{Tr}\{\Lambda_{m,k}^{B^n}\rho_{x^n(m,k)}^{B^n}\} \geq 1-\epsilon, \tag{23.12}$$

while Eve obtains asymptotically zero information about Alice's message m:

$$\forall m \in \mathcal{M} : \left\| \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}} \omega_{x^n(m,k)}^{E^n} - \omega^{\otimes n} \right\|_1 \leq \epsilon, \tag{23.13}$$

where ω is Eve's expected density operator:

$$\omega \equiv \sum_x p_X(x)\omega_x^E. \tag{23.14}$$

The above statements hold true for all $\epsilon > 0$ and sufficiently large n as long as

$$|\mathcal{M}| \approx 2^{n[I(X;B)-I(X;E)]}, \tag{23.15}$$

$$|\mathcal{K}| \approx 2^{nI(X;E)}. \tag{23.16}$$

We can now construct a coherent version of the above code that is good for quantum data transmission. First, suppose that there is some density operator with the following spectral decomposition:

$$\rho^{A'} \equiv \sum_x p_X(x)|\psi_x\rangle\langle\psi_x|^{A'}. \tag{23.17}$$

Now suppose that the channel $\mathcal{N}^{A'\to B}$ has an isometric extension $U_{\mathcal{N}}^{A'\to BE}$, so that inputting $|\psi_x\rangle^{A'}$ leads to the following state shared between Bob and Eve:

$$|\psi_x\rangle^{BE} \equiv U_{\mathcal{N}}^{A'\to BE}|\psi_x\rangle^{A'}. \tag{23.18}$$

From the direct coding part of Theorem 22.2.1, we know that there exists a private classical code $\{x^n(m,k)\}_{m\in\mathcal{M},k\in\mathcal{K}}$ with the properties in (23.12-23.13) and with rate

$$I(X;B)_\sigma - I(X;E)_\sigma, \tag{23.19}$$

where σ^{XBE} is a classical-quantum state of the following form:

$$\sigma^{XBE} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes |\psi_x\rangle\langle\psi_x|^{BE}. \tag{23.20}$$

The following identity demonstrates that the private information in (23.19) is equal to the coherent information for the particular state σ^{XBE} above:

$$\begin{aligned} I(X;B)_\sigma - I(X;E)_\sigma &= H(B)_\sigma - H(B|X)_\sigma - H(E)_\sigma + H(E|X)_\sigma & (23.21)\\ &= H(B)_\sigma - H(B|X)_\sigma - H(E)_\sigma + H(B|X)_\sigma & (23.22)\\ &= H(B)_\sigma - H(E)_\sigma. & (23.23) \end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first equality follows from the identity $I(C;D) = H(D) - H(D|C)$. The second equality follows because the state on systems BE is pure when conditioned on the classical variable X. Observe that the last expression is a function solely of the density operator ρ in (23.17), and it is also equal to the coherent information of the channel for the particular input state ρ (see Exercise 11.5.2).

Now we show how to construct a quantum code achieving the coherent information rate $H(B)_\sigma - H(E)_\sigma$ by making a coherent version of the above private classical code. Suppose that Alice shares a state $|\varphi\rangle^{RA_1}$ with a reference system R, where

$$|\varphi\rangle^{RA_1} \equiv \sum_{l,m\in\mathcal{M}} \alpha_{l,m}|l\rangle^R|m\rangle^{A_1}, \tag{23.24}$$

and $\{|l\rangle^R\}$ is some orthonormal basis for R while $\{|m\rangle^{A_1}\}$ is some orthonormal basis for A_1. Also, we set $|\mathcal{M}| \approx 2^{n\left[H(B)_\sigma - H(E)_\sigma\right]}$. We would like for Alice and Bob to execute a quantum communication protocol such that Bob can reconstruct Alice's share of the above state on his system with Alice no longer entangled with the reference (we would like for the final state to be approximately close to $|\varphi\rangle^{RA_1}$ where Bob is holding the A_1 system). To this end, Alice creates a quantum codebook $\{|\phi_m\rangle^{A'^n}\}_{m\in\mathcal{M}}$ with quantum codewords:

$$|\phi_m\rangle^{A'^n} \equiv \frac{1}{\sqrt{|\mathcal{K}|}} \sum_{k\in\mathcal{K}} e^{i\gamma_{m,k}}|\psi_{x^n(m,k)}\rangle^{A'^n}, \tag{23.25}$$

where the states $|\psi_{x^n(m,k)}\rangle^{A'^n}$ are the n^{th} extensions of the states arising from the spectral decomposition in (23.17), the classical sequences $x^n(m,k)$ are from the codebook for private classical communication, and we specify how to choose the phases $\gamma_{m,k}$ later. All the states $|\psi_{x^n(m,k)}\rangle^{A'^n}$ are orthonormal because they are picked from the spectral decomposition in (23.17) and the expurgation from Section 22.3.4 guarantees that they are distinct (otherwise, they would not be good codewords!). The fact that the states $|\psi_{x^n(m,k)}\rangle^{A'^n}$ are orthonormal implies that the quantum codewords $|\phi_m\rangle^{A'^n}$ are also orthonormal.

Alice's first action is to coherently copy the value of m in the A_1 register to another register A_2, so that the state in (23.24) becomes

$$\sum_{l,m} \alpha_{l,m}|l\rangle^R|m\rangle^{A_1}|m\rangle^{A_2}. \tag{23.26}$$

Alice then performs some isometric encoding from A_2 to A'^n that takes the above unencoded state to the following encoded state:

$$\sum_{l,m} \alpha_{l,m}|l\rangle^R|m\rangle^{A_1}|\phi_m\rangle^{A'^n}, \tag{23.27}$$

where each $|\phi_m\rangle^{A'^n}$ is a quantum codeword of the form in (23.25). Alice transmits the systems A'^n through many uses of the quantum channel, leading to the following state shared between the reference, Alice, Bob, and Eve:

$$\sum_{l,m} \alpha_{l,m}|l\rangle^R|m\rangle^{A_1}|\phi_m\rangle^{B^nE^n}, \tag{23.28}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $|\phi_m\rangle^{B^nE^n}$ is defined from (23.18) and (23.25). Recall from (23.12) that Bob can detect the message m and the variable k in the private classical code with high probability:

$$\forall m, k : \text{Tr}\{\Lambda^{B^n}_{m,k}\psi^{B^n}_{x^n(m,k)}\} \geq 1 - \epsilon. \tag{23.29}$$

So Bob instead constructs a coherent version of this POVM:

$$\sum_{m\in\mathcal{M},k\in\mathcal{K}} \sqrt{\Lambda_{m,k}}^{B^n} \otimes |m\rangle^{B_1}|k\rangle^{B_2}. \tag{23.30}$$

He then performs this coherent POVM, resulting in the state

$$\sum_{\substack{m'\in\mathcal{M},\\ k'\in\mathcal{K}}} \sum_{l,m} \sum_{k\in\mathcal{K}} \frac{1}{\sqrt{|\mathcal{K}|}} \alpha_{l,m}|l\rangle^R|m\rangle^{A_1} \sqrt{\Lambda_{m',k'}}^{B^n} e^{i\gamma_{k,m}}|\psi_{x^n(m,k)}\rangle^{B^nE^n}|m',k'\rangle^{B_1B_2}. \tag{23.31}$$

We would like for the above state to be close in trace distance to the following state:

$$\sum_{l,m} \sum_{k\in\mathcal{K}} \frac{1}{\sqrt{|\mathcal{K}|}} \alpha_{l,m}|l\rangle^R|m\rangle^{A_1} e^{i\delta_{m,k}}|\psi_{x^n(m,k)}\rangle^{B^nE^n}|m\rangle^{B_1}|k\rangle^{B_2}, \tag{23.32}$$

where $\delta_{m,k}$ are some phases that we will specify shortly. To this end, consider that the sets $\{|\chi_{m,k}\rangle^{B^nE^nB_1B_2}\}_{m,k}$ and $\{|\varphi_{m,k}\rangle^{B^nE^nB_1B_2}\}_{m,k}$ form orthonormal bases, where

$$|\chi_{m,k}\rangle^{B^nE^nB_1B_2} \equiv |\psi_{x^n(m,k)}\rangle^{B^nE^n}|m\rangle^{B_1}|k\rangle^{B_2}, \tag{23.33}$$

$$|\varphi_{m,k}\rangle^{B^nE^nB_1B_2} \equiv \sum_{m'\in\mathcal{M},k'\in\mathcal{K}} \sqrt{\Lambda_{m',k'}}^{B^n}|\psi_{x^n(m,k)}\rangle^{B^nE^n}|m'\rangle^{B_1}|k'\rangle^{B_2}. \tag{23.34}$$

Also, consider that the overlap between corresponding states in the different bases is high:

$$\begin{aligned}
&\langle\chi_{m,k}|\varphi_{m,k}\rangle \\
&= \langle\psi_{x^n(m,k)}|^{B^nE^n}\langle m|^{B_1}\langle k|^{B_2} \sum_{m'\in\mathcal{M},k'\in\mathcal{K}} \sqrt{\Lambda_{m',k'}}^{B^n}|\psi_{x^n(m,k)}\rangle^{B^nE^n}|m'\rangle^{B_1}|k'\rangle^{B_2} && (23.35)\\
&= \sum_{m'\in\mathcal{M},k'\in\mathcal{K}} \langle\psi_{x^n(m,k)}|^{B^nE^n}\sqrt{\Lambda_{m',k'}}^{B^n}|\psi_{x^n(m,k)}\rangle^{B^nE^n}\langle m|m'\rangle^{B_1}\langle k|k'\rangle^{B_2} && (23.36)\\
&= \langle\psi_{x^n(m,k)}|^{B^nE^n}\sqrt{\Lambda_{m,k}}^{B^n}|\psi_{x^n(m,k)}\rangle^{B^nE^n} && (23.37)\\
&\geq \langle\psi_{x^n(m,k)}|^{B^nE^n}\Lambda^{B^n}_{m,k}|\psi_{x^n(m,k)}\rangle^{B^nE^n} && (23.38)\\
&= \text{Tr}\{\Lambda^{B^n}_{m,k}\psi^{B^n}_{x^n(m,k)}\} && (23.39)\\
&\geq 1-\epsilon. && (23.40)
\end{aligned}$$

where the first inequality follows from the fact that $\sqrt{\Lambda_{m,k}}^{B^n} \geq \Lambda^{B^n}_{m,k}$ for $\Lambda^{B^n}_{m,k} \leq I$ and the second inequality follows from (23.29). By applying Lemma A.0.4 from Appendix A, we know that there exist phases $\gamma_{m,k}$ and $\delta_{m,k}$ such that

$$\langle\chi_m|\varphi_m\rangle \geq 1-\epsilon, \tag{23.41}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where

$$|\chi_m\rangle^{B^nE^nB_1B_2} \equiv \frac{1}{\sqrt{|\mathcal{K}|}}\sum_k e^{i\delta_{m,k}}|\chi_{m,k}\rangle^{B^nE^nB_1B_2}, \tag{23.42}$$

$$|\varphi_m\rangle^{B^nE^nB_1B_2} \equiv \frac{1}{\sqrt{|\mathcal{K}|}}\sum_k e^{i\gamma_{m,k}}|\varphi_{m,k}\rangle^{B^nE^nB_1B_2}. \tag{23.43}$$

So we choose the phases in a way such that the above inequality holds. We can then apply the above result to show that the state in (23.31) has high fidelity with the state in (23.32):

$$\left(\sum_{l,m}\alpha^*_{l,m}\langle l|^R\langle m|^{A_1}\langle\chi_m|^{B^nE^nB_1B_2}\right)\left(\sum_{l',m'}\alpha_{l',m'}|l'\rangle^R|m'\rangle^{A_1}|\varphi_{m'}\rangle^{B^nE^nB_1B_2}\right)$$

$$= \sum_{l,m,l',m'}\alpha^*_{l,m}\alpha_{l',m'}\langle l|l'\rangle^R\langle m|m'\rangle^{A_1}\langle\chi_m|\varphi_{m'}\rangle^{B^nE^nB_1B_2} \tag{23.44}$$

$$= \sum_{l,m}|\alpha_{l,m}|^2\langle\chi_m|\varphi_m\rangle^{B^nE^nB_1B_2} \tag{23.45}$$

$$\geq 1-\epsilon. \tag{23.46}$$

Thus, the state resulting after Bob performs the coherent POVM is close in trace distance to the following state:

$$\sum_{l,m\in\mathcal{M}}\alpha_{l,m}|l\rangle^R|m\rangle^{A_1}\frac{1}{\sqrt{|\mathcal{K}|}}\sum_{k\in\mathcal{K}}e^{i\delta_{m,k}}|\psi_{x^n(m,k)}\rangle^{B^nE^n}|m\rangle^{B_1}|k\rangle^{B_2}$$

$$= \sum_{l,m\in\mathcal{M}}\alpha_{l,m}|l\rangle^R|m\rangle^{A_1}|\widetilde{\phi}_m\rangle^{B^nE^nB_2}|m\rangle^{B_1}, \tag{23.47}$$

where

$$|\widetilde{\phi}_m\rangle^{B^nE^nB_2} \equiv \frac{1}{\sqrt{|\mathcal{K}|}}\sum_{k\in\mathcal{K}}e^{i\delta_{m,k}}|\psi_{x^n(m,k)}\rangle^{B^nE^n}|k\rangle^{B_2}. \tag{23.48}$$

Consider the state of Eve for a particular value of m:

$$\widetilde{\phi}^{E^n}_m = \mathrm{Tr}_{B^nB_2}\left\{|\widetilde{\phi}_m\rangle\langle\widetilde{\phi}_m|^{B^nE^nB_2}\right\} \tag{23.49}$$

$$= \mathrm{Tr}_{B^nB_2}\left\{\sum_{k,k'\in\mathcal{K}}\frac{1}{|\mathcal{K}|}e^{i(\delta_{m,k'}-\delta_{m,k})}|\psi_{x^n(m,k)}\rangle\langle\psi_{x^n(m,k')}|^{B^nE^n}\otimes|k\rangle\langle k'|^{B_2}\right\} \tag{23.50}$$

$$= \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\psi^{E^n}_{x^n(m,k)}. \tag{23.51}$$

We are now in a position to apply the second property of the private classical code. Recall from the privacy condition in (23.13) that Eve's state is guaranteed to be ϵ-close in trace distance to the tensor power state $\left[(\mathcal{N}^c)^{A'\to E}(\rho)\right]^{\otimes n}$, where $(\mathcal{N}^c)^{A'\to E}$ is the complementary

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

channel and ρ is the density operator in (23.17). Let $|\theta_{\mathcal{N}^c(\rho)}\rangle^{E^nB_3}$ be some purification of this tensor power state. By Uhlmann's theorem and the relation between trace distance and fidelity (see Definition 9.2.3 and Theorem 9.3.1), there is some isometry $U_m^{B^nB_2\to B_3}$ for each value of m such that the following states are $2\sqrt{\epsilon}$-close in trace distance (see Exercise 9.2.7):

$$U_m^{B^nB_2\to B_3}|\widetilde{\phi}_m\rangle^{B^nE^nB_2} \overset{2\sqrt{\epsilon}}{\approx} |\theta_{\mathcal{N}^c(\rho)}\rangle^{E^nB_3}. \tag{23.52}$$

Bob's next step is to perform the following controlled isometry on his systems B^n, B_1, and B_2:

$$\sum_m |m\rangle\langle m|^{B_1} \otimes U_m^{B^nB_2\to B_3}, \tag{23.53}$$

leading to a state that is close in trace distance to the following state:

$$\left(\sum_{l,m\in\mathcal{M}} \alpha_{l,m}|l\rangle^R|m\rangle^{A_1}|m\rangle^{B_1}\right) \otimes |\theta_{\mathcal{N}^c(\rho)}\rangle^{E^nB_3}. \tag{23.54}$$

At this point, the key observation is that the state on E^nB_3 is effectively decoupled from the state on systems R, A_1, and B_1, so that Bob can just throw away his system B_3. Thus, they have successfully implemented an approximate coherent channel from system A_1 to A_1B_1.

We now allow for Alice to communicate classical information to Bob in order for them to implement a quantum communication channel rather than just a mere coherent channel (in a moment we argue that this free forward classical communication is not necessary). Alice performs a Fourier transform on the register A_1, leading to the following state:

$$\frac{1}{\sqrt{d_{A_1}}} \sum_{l,m,j\in\mathcal{M}} \alpha_{l,m} \exp\{2\pi imj/d_{A_1}\}|l\rangle^R|j\rangle^{A_1}|m\rangle^{B_1}. \tag{23.55}$$

She then measures register A_1 in the computational basis, leading to some outcome j and the following post-measurement state:

$$\left(\sum_{l,m\in\mathcal{M}} \alpha_{l,m} \exp\{2\pi imj/d_{A_1}\}|l\rangle^R|m\rangle^{B_1}\right) \otimes |j\rangle^{A_1}. \tag{23.56}$$

She sends Bob the outcome j of her measurement over a classical channel, and the protocol ends with Bob performing the following unitary

$$Z^\dagger(j)|m\rangle^{B_1} = \exp\{-2\pi imj/d_{A_1}\}|m\rangle^{B_1}, \tag{23.57}$$

leaving the desired state on the reference and Bob's system B_1:

$$\sum_{l,m\in\mathcal{M}} \alpha_{l,m}|l\rangle^R|m\rangle^{B_1}. \tag{23.58}$$

All of the errors accumulated in the above protocol are some finite sum of ϵ terms, and applying the triangle inequality several times implies that the actual state is close to the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

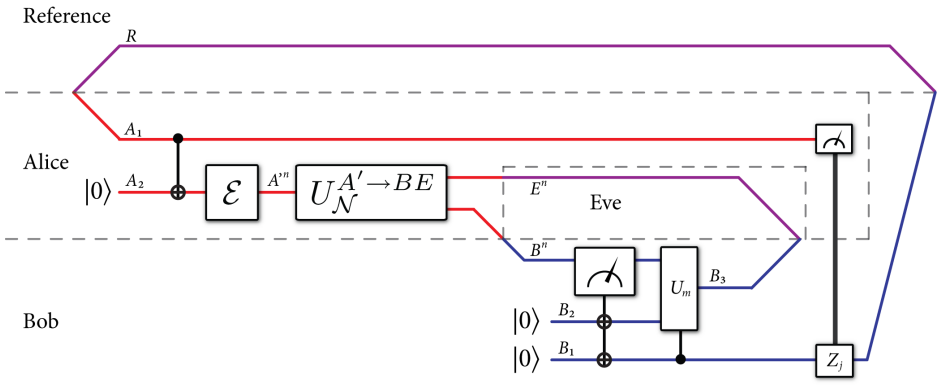

Figure 23.2: All of the steps in the protocol for quantum communication. Alice and Bob's goal is to communicate as much quantum information as they can while making sure that Eve's state is independent of what Alice is trying to communicate to Bob. The figure depicts the series of controlled unitaries that Alice and Bob perform and the final measurement and classical communication that enables quantum communication from Alice to Bob at the coherent information rate.

desired state in the asymptotic limit of large block length. Figure 23.2 depicts all of the steps in this protocol for quantum communication.

We now argue that the classical communication is not necessary—there exists a scheme that does not require the use of this forward classical channel. After reviewing the above protocol and glancing at Figure 23.2, we realize that Alice's encoder is a quantum instrument of the following form:

$$\mathcal{E}(\rho) \equiv \sum_j \mathcal{E}_j(\rho) \otimes |j\rangle\langle j|. \tag{23.59}$$

Each map $\mathcal{E}_j(\rho)$ is a trace-reducing map of the following form:

$$\mathcal{E}_j\left(\psi^{RA_1}\right) = \langle j|^{A_1} F^{A_1} \left(\sum_{m'} |\phi_{m'}\rangle\langle m'|^{A_2}\right)\left(\sum_m |m\rangle\langle m|^{A_1} \otimes |m\rangle^{A_2}\right)|\psi\rangle^{RA_1}, \tag{23.60}$$

where $\sum_m |m\rangle\langle m|^{A_1} \otimes |m\rangle^{A_2}$ is Alice's coherent copier in (23.26), $\sum_{m'} |\phi_{m'}\rangle\langle m'|^{A_2}$ is her quantum encoder in (23.27), F^{A_1} is the Fourier transform, and $\langle j|^{A_1}$ represents the projection onto a particular measurement outcome j. We can simplify the above map as follows:

$$\begin{aligned}
\mathcal{E}_j\left(\psi^{RA_1}\right) &= \langle j|^{A_1} \sum_{m'} |\widetilde{m}'\rangle\langle m'|^{A_1} \left(\sum_m |m\rangle\langle m|^{A_1} \otimes |\phi_m\rangle^{A_2}\right)|\psi\rangle^{RA_1} && (23.61)\\
&= \langle j|^{A_1} \sum_m |\widetilde{m}\rangle\langle m|^{A_1} \otimes |\phi_m\rangle^{A_2}|\psi\rangle^{RA_1} && (23.62)\\
&= \left(\frac{1}{\sqrt{|\mathcal{M}|}} \sum_m e^{i2\pi mj/|\mathcal{M}|} |\phi_m\rangle^{A_2}\langle m|^{A_1}\right)|\psi\rangle^{RA_1} && (23.63)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

It follows that the trace of each $\mathcal{E}_j$ is uniform and independent of the input state $|\psi\rangle^{RA_1}$:

$$\text{Tr}\{\mathcal{E}_j(\psi^{RA_1})\} = \frac{1}{|\mathcal{M}|}. \tag{23.64}$$

Observe that multiplying the map in (23.63) by $\sqrt{|\mathcal{M}|}$ gives a proper isometry that could suffice as an encoding. Let $\mathcal{E}'_j$ denote the rescaled isometry. Corresponding to each encoder is a decoding map $\mathcal{D}_j$ consisting of Bob's coherent measurement in (23.30), his decoupler in (23.53), and his phase shifter in (23.57). We can thus represent the state output from our classically-coordinated protocol as follows:

$$\sum_j \mathcal{D}_j(\mathcal{N}^{\otimes n}(\mathcal{E}_j(\varphi^{RA_1}))). \tag{23.65}$$

From the analysis in the preceding paragraphs, we know that the trace distance between the ideal state and the actual state is small for the classically-coordinated scheme:

$$\left\| \sum_j \mathcal{D}_j(\mathcal{N}^{\otimes n}(\mathcal{E}_j(\varphi^{RA_1}))) - \varphi^{RA_1} \right\|_1 \leq \epsilon', \tag{23.66}$$

where ϵ' is some arbitrarily small positive number. Thus, the fidelity between these two states is high:

$$F\left(\sum_j \mathcal{D}_j(\mathcal{N}^{\otimes n}(\mathcal{E}_j(\varphi^{RA_1}))), \varphi^{RA_1} \right) \geq 1 - \epsilon'. \tag{23.67}$$

But we can rewrite the fidelity as follows:

$$\begin{aligned}
&F\left(\sum_j \mathcal{D}_j(\mathcal{N}^{\otimes n}(\mathcal{E}_j(\varphi^{RA_1}))), \varphi^{RA_1} \right) \\
&= \langle\varphi|^{RA_1} \sum_j \mathcal{D}_j(\mathcal{N}^{\otimes n}(\mathcal{E}_j(\varphi^{RA_1})))|\varphi\rangle^{RA_1} && (23.68)\\
&= \sum_j \langle\varphi|^{RA_1} \mathcal{D}_j(\mathcal{N}^{\otimes n}(\mathcal{E}_j(\varphi^{RA_1})))|\varphi\rangle^{RA_1} && (23.69)\\
&= \sum_j \frac{1}{|\mathcal{M}|} \left[\langle\varphi|^{RA_1} \mathcal{D}_j(\mathcal{N}^{\otimes n}(\mathcal{E}'_j(\varphi^{RA_1})))|\varphi\rangle^{RA_1} \right] && (23.70)\\
&\geq 1 - \epsilon', && (23.71)
\end{aligned}$$

implying that at least one of the encoder-decoder pairs $(\mathcal{E}'_j, \mathcal{D}_j)$ has asymptotically high fidelity. Thus, Alice and Bob simply agree beforehand to use a scheme $(\mathcal{E}'_j, \mathcal{D}_j)$ with high fidelity, obviating the need for the forward classical communication channel.

The protocol given here achieves communication at the coherent information rate. In order to achieve the regularized coherent information rate in the statement of the theorem, Alice and Bob apply the same protocol to the superchannel $(\mathcal{N}^{A'\to B})^{\otimes k}$ instead of the channel $\mathcal{N}^{A'\to B}$.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

23.5 Converse Theorem

This section proves the converse part of the quantum capacity theorem, demonstrating that the regularized coherent information is an upper bound on the quantum capacity of any quantum channel. For the class of degradable channels, the coherent information itself is an upper bound on the quantum capacity—this demonstrates that we completely understand the quantum data transmission capabilities of these channels.

For this converse proof, we assume that Alice is trying to generate entanglement with Bob. The capacity for this task is an upper bound on the capacity for quantum data transmission because we can always use a noiseless quantum channel to establish entanglement. We also allow Alice free forward classical communication to Bob, and we demonstrate that this resource cannot increase the quantum capacity (essentially because the coherent information is convex). In a protocol for entanglement generation, Alice begins by preparing the maximally entangled state Φ^{AA_1} of Schmidt rank 2^{nQ} in her local laboratory, where Q is the rate of this entangled state. She performs some encoding operation $\mathcal{E}^{A_1 \to A'^n M}$ that outputs many systems A'^n and a classical register M. She then inputs the systems A'^n to many independent uses of a noisy quantum channel $\mathcal{N}^{A' \to B}$, resulting in the state

$$\omega^{AMB^n} \equiv \mathcal{N}^{A'^n \to B^n}(\mathcal{E}^{A_1 \to A'^n M}(\Phi^{AA_1})), \tag{23.72}$$

where $\mathcal{N}^{A'^n \to B^n} \equiv (\mathcal{N}^{A' \to B})^{\otimes n}$. Bob takes the outputs B^n of the channels and the classical register M and performs some decoding operation $\mathcal{D}^{B^n M \to B_1}$, resulting in the state

$$(\omega')^{AB_1} \equiv \mathcal{D}^{B^n M \to B_1}(\omega^{AMB^n}). \tag{23.73}$$

If the protocol is any good for entanglement generation, then the following condition should hold

$$\left\| (\omega')^{AB_1} - \Phi^{AB_1} \right\|_1 \leq \epsilon, \tag{23.74}$$

where ϵ is some arbitrarily small positive number.

The converse proof then proceeds in the following steps:

$$\begin{aligned} nQ &= I(A\rangle B_1)_{\Phi} && (23.75)\\ &\leq I(A\rangle B_1)_{\omega'} + n\epsilon' && (23.76)\\ &\leq I(A\rangle B^n M)_{\omega} + n\epsilon'. && (23.77) \end{aligned}$$

The first equality follows because the coherent information of a maximally entangled state is equal to the logarithm of the dimension of one of its systems. The first inequality follows from an application of the Alicki-Fannes' inequality to the condition in (23.74), with $\epsilon' \equiv 4\epsilon Q + 2H_2(\epsilon)/n$. The second inequality follows from quantum data processing. Now consider that the state ω^{AMB^n} is a classical-quantum state of the following form:

$$\omega^{AMB^n} \equiv \sum_m p_M(m)|m\rangle\langle m|^M \otimes \mathcal{N}^{A'^n \to B^n}(\rho_m^{AA'^n}). \tag{23.78}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can then perform a spectral decomposition of each state ρ_m as follows:

$$\rho_m^{AA'^n} = \sum_l p_{L|M}(l|m)|\phi_{l,m}\rangle\langle\phi_{l,m}|^{AA'^n}, \tag{23.79}$$

and augment the above state as follows:

$$\omega^{AMLB^n} \equiv \sum_{m,l} p_M(m)p_{L|M}(l|m)|m\rangle\langle m|^M \otimes |l\rangle\langle l|^L \otimes \mathcal{N}^{A'^n\to B^n}(\phi_{l,m}^{AA'^n}), \tag{23.80}$$

so that $\omega^{AMB^n} = \mathrm{Tr}_L\{\omega^{AMLB^n}\}$. We continue with bounding the rate Q:

$$\begin{aligned} I(A\rangle B^n M)_\omega + n\epsilon' &\leq I(A\rangle B^n ML)_\omega + n\epsilon' && (23.81)\\ &= \sum_{m,l} p_M(m)p_{L|M}(l|m) I(A\rangle B^n)_{\mathcal{N}^{A'^n\to B^n}(\phi_{l,m}^{AA'^n})} + n\epsilon' && (23.82)\\ &\leq I(A\rangle B^n)_{\mathcal{N}^{A'^n\to B^n}((\phi^*)_{l,m}^{AA'^n})} + n\epsilon' && (23.83)\\ &\leq Q\big(\mathcal{N}^{\otimes n}\big) + n\epsilon'. && (23.84)\end{aligned}$$

The first inequality follows from the quantum data processing inequality. The first equality follows because the registers M and L are both classical, and we can apply the result of Exercise 11.5.5. The second inequality follows because the expectation is always less than the maximal value (where we define ϕ^* to be the state that achieves this maximum). The final inequality follows from the definition of the channel coherent information as the maximum of the coherent information over all pure, bipartite inputs. This concludes the proof of the converse part of the quantum capacity theorem.

There are a few comments we should make regarding the converse theorem. First, we see that classical communication cannot improve quantum capacity because the coherent information is convex. We could obtain the same upper bound on quantum capacity even if there were no classical communication. Second, it is sufficient to consider isometric encoders for quantum communication—that is, it is not necessary to exploit general noisy CPTP maps at the encoder. This makes sense intuitively because it would seem odd if noisy encodings could help in the noiseless transmission of quantum data. Our augmented state in (23.80) and the subsequent development reveals that this is so (again because the coherent information is convex).

We can significantly strengthen the statement of the quantum capacity theorem for the class of degradable quantum channels because the following inequality holds for them:

$$Q\big(\mathcal{N}^{\otimes n}\big) \leq nQ(\mathcal{N}). \tag{23.85}$$

This inequality follows from the additivity of coherent information for degradable channels (Theorem 12.5.1). Also, the task of optimizing the coherent information for these channels is straightforward because it is a concave function of the input density operator (Theorem 12.5.2) and the set of density operators is convex.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

23.6 Example Channels

We now show how to calculate the quantum capacity for two exemplary channels: the quantum erasure channel and the amplitude damping channel. Both of these channels are degradable, simplifying the calculation of their quantum capacities.

23.6.1 The Quantum Erasure Channel

Recall that the quantum erasure channel acts as follows on an input density operator $\rho^{A'}$:

$$\rho^{A'} \rightarrow (1-\epsilon)\rho^{B} + \epsilon |e\rangle\langle e|^{B}, \tag{23.86}$$

where ϵ is the erasure probability and $|e\rangle^{B}$ is an erasure state that is orthogonal to the support of any input state ρ.

Proposition 23.6.1. *The quantum capacity of a quantum erasure channel with erasure probability ϵ is*

$$(1-2\epsilon)\log d_A, \tag{23.87}$$

where d_A is the dimension of the input system.

Proof. To determine the quantum capacity of this channel, we need to compute its coherent information, and we can do so in a similar way as we did in Proposition 20.6.1. So, consider that sending half of a pure, bipartite state $\phi^{AA'}$ through the channel produces the output

$$\sigma^{AB} \equiv (1-\epsilon)\phi^{AB} + \epsilon \phi^{A} \otimes |e\rangle\langle e|^{B}. \tag{23.88}$$

Recall that Bob can apply the following isometry $U^{B\rightarrow BX}$ to his state:

$$U^{B\rightarrow BX} \equiv \Pi^{B} \otimes |0\rangle^{X} + |e\rangle\langle e|^{B} \otimes |1\rangle^{X}, \tag{23.89}$$

where Π^{B} is a projector onto the support of the input state (for qubits, it would be just $|0\rangle\langle 0| + |1\rangle\langle 1|$). Applying this isometry leads to a state σ^{ABX} where

$$\begin{aligned} \sigma^{ABX} &\equiv U^{B\rightarrow BX}\sigma^{AB}\left(U^{B\rightarrow BX}\right)^{\dagger} && (23.90)\\ &= (1-\epsilon)\phi^{AB} \otimes |0\rangle\langle 0|^{X} + \epsilon \phi^{A} \otimes |e\rangle\langle e|^{B} \otimes |1\rangle\langle 1|^{X}. && (23.91) \end{aligned}$$

The coherent information $I(A\rangle BX)_{\sigma}$ is equal to $I(A\rangle B)_{\sigma}$ because entropies do not change

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

under the isometry $U^{B\to BX}$. We now calculate $I(A\rangle BX)_\sigma$:

$$\begin{aligned}
I(A\rangle BX)_\sigma &= H(BX)_\sigma - H(ABX)_\sigma && (23.92)\\
&= H(B|X)_\sigma - H(AB|X)_\sigma && (23.93)\\
&= (1-\epsilon)\Big[H(B)_\phi - H(AB)_\phi\Big] \\
&\quad + \epsilon\Big[H(B)_{|e\rangle} - H(AB)_{\phi^A\otimes|e\rangle\langle e|}\Big] && (23.94)\\
&= (1-\epsilon)H(B)_\phi - \epsilon\Big[H(A)_\phi + H(B)_{|e\rangle}\Big] && (23.95)\\
&= (1-2\epsilon)H(A)_\phi && (23.96)\\
&\leq (1-2\epsilon)\log d_A. && (23.97)
\end{aligned}$$

The first equality follows by the definition of coherent information. The second equality follows from $\phi^A = \mathrm{Tr}_{BX}\{\sigma^{ABX}\}$, from the chain rule of entropy, and by canceling $H(X)$ on both sides. The third equality follows because the X register is a classical register, indicating whether the erasure occurs. The fourth equality follows because $H(AB)_\phi = 0$, $H(B)_{|e\rangle} = 0$, and $H(AB)_{\phi^A\otimes|e\rangle\langle e|} = H(A)_\phi + H(B)_{|e\rangle}$. The fifth equality follows again because $H(B)_{|e\rangle} = 0$, by collecting terms, and because $H(A)_\phi = H(B)_\phi$ (ϕ^{AB} is a pure bipartite state). The final inequality follows because the entropy of a state on system A is never greater than the logarithm of the dimension of A. We can conclude that the maximally entangled state $\Phi^{AA'}$ achieves the entanglement-assisted classical capacity of the quantum erasure channel because $H(A)_\Phi = \log d_A$. □

23.6.2 The Amplitude Damping Channel

We now compute the quantum capacity of the amplitude damping channel $\mathcal{N}_{\mathrm{AD}}$. Recall that this channel acts as follows on an input qubit in state ρ:

$$\mathcal{N}_{\mathrm{AD}}(\rho) = A_0\rho A_0^\dagger + A_1\rho A_1^\dagger, \tag{23.98}$$

where

$$A_0 \equiv |0\rangle\langle 0| + \sqrt{1-\gamma}|1\rangle\langle 1|, \qquad A_1 \equiv \sqrt{\gamma}|0\rangle\langle 1|. \tag{23.99}$$

The development here is similar to development in the proof of Proposition 20.6.2.

Proposition 23.6.2. *The quantum capacity of an amplitude damping channel with damping parameter γ is*

$$Q(\mathcal{N}_{AD}) = \max_{p\in[0,1]} H_2((1-\gamma)p) - H_2(\gamma p), \tag{23.100}$$

whenever $\gamma \leq 1/2$. Otherwise, the quantum capacity is zero. Recall that $H_2(x)$ is the binary entropy function.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Proof. Suppose that a matrix representation of the input qubit density operator ρ in the computational basis is

$$\rho = \begin{bmatrix} 1-p & \eta^* \\ \eta & p \end{bmatrix}. \tag{23.101}$$

One can readily verify that the density operator for Bob has the following matrix representation:

$$\mathcal{N}_{\text{AD}}(\rho) = \begin{bmatrix} 1-(1-\gamma)p & \sqrt{1-\gamma}\eta^* \\ \sqrt{1-\gamma}\eta & (1-\gamma)p \end{bmatrix}, \tag{23.102}$$

and by calculating the elements $\text{Tr}\{A_i \rho A_j^\dagger\}|i\rangle\langle j|$, we can obtain a matrix representation for Eve's density operator:

$$\mathcal{N}^c_{\text{AD}}(\rho) = \begin{bmatrix} 1-\gamma p & \sqrt{\gamma}\eta^* \\ \sqrt{\gamma}\eta & \gamma p \end{bmatrix}, \tag{23.103}$$

where $\mathcal{N}^c_{\text{AD}}$ is the complementary channel to Eve. By comparing (23.102) and (23.103), we can see that the channel to Eve is an amplitude damping channel with damping parameter $1-\gamma$. The quantum capacity of $\mathcal{N}_{\text{AD}}$ is equal to its coherent information:

$$Q(\mathcal{N}_{\text{AD}}) = \max_{\phi^{AA'}} I(A\rangle B)_\sigma, \tag{23.104}$$

where $\phi^{AA'}$ is some pure bipartite input state and $\sigma^{AB} = \mathcal{N}_{\text{AD}}(\phi^{AA'})$. We need to determine the input density operator that maximizes the above formula as a function of γ. So far, the optimization depends on three parameters: p, $\text{Re}\{\eta\}$, and $\text{Im}\{\eta\}$. We can show that it is sufficient to consider an optimization over only p with $\eta = 0$. The formula in (23.104) also has the following form:

$$Q(\mathcal{N}_{\text{AD}}) = \max_{\rho} [H(\mathcal{N}_{\text{AD}}(\rho)) - H(\mathcal{N}^c_{\text{AD}}(\rho))], \tag{23.105}$$

because

$$\begin{aligned} I(A\rangle B)_\sigma &= H(B)_\sigma - H(AB)_\sigma && (23.106)\\ &= H(\mathcal{N}_{\text{AD}}(\rho)) - H(E)_\sigma && (23.107)\\ &= H(\mathcal{N}_{\text{AD}}(\rho)) - H(\mathcal{N}^c_{\text{AD}}(\rho)) && (23.108)\\ &\equiv I_{\text{coh}}(\rho, \mathcal{N}_{\text{AD}}). && (23.109) \end{aligned}$$

The two entropies in (23.105) depend only on the eigenvalues of the two density operators in (23.102-23.103), respectively, which are as follows:

$$\frac{1}{2}\left(1 \pm \sqrt{(1-2(1-\gamma)p)^2 + 4|\eta|^2(1-\gamma)}\right), \tag{23.110}$$

$$\frac{1}{2}\left(1 \pm \sqrt{(1-2\gamma p)^2 + 4|\eta|^2\gamma}\right). \tag{23.111}$$

The above eigenvalues are in the order of Bob and Eve. All of the above eigenvalues have a similar form, and their dependence on η is only through its magnitude. Thus, it suffices to

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

consider $\eta \in \mathbb{R}$ (this eliminates one parameter). Next, the eigenvalues do not change if we flip the sign of η (this is equivalent to rotating the original state ρ by Z, to $Z\rho Z$), and thus, the coherent information does not change as well:

$$I_{\text{coh}}(\rho, \mathcal{N}_{\text{AD}}) = I_{\text{coh}}(Z\rho Z, \mathcal{N}_{\text{AD}}). \tag{23.112}$$

By the above relation and concavity of coherent information in the input density operator for degradable channels (Theorem 12.5.2), the following inequality holds

$$\begin{aligned} I_{\text{coh}}(\rho, \mathcal{N}_{\text{AD}}) &= \frac{1}{2}[I_{\text{coh}}(\rho, \mathcal{N}_{\text{AD}}) + I_{\text{coh}}(Z\rho Z, \mathcal{N}_{\text{AD}})] & (23.113)\\ &\leq I_{\text{coh}}(\frac{1}{2}(\rho + Z\rho Z), \mathcal{N}_{\text{AD}}) & (23.114)\\ &= I_{\text{coh}}(\overline{\Delta}(\rho), \mathcal{N}_{\text{AD}}), & (23.115) \end{aligned}$$

where $\overline{\Delta}$ is a completely dephasing channel in the computational basis. This demonstrates that it is sufficient to consider diagonal density operators ρ when optimizing the coherent information. Thus, the eigenvalues in (23.110-23.111) respectively become

$$\begin{aligned} &\{(1-\gamma)p, 1-(1-\gamma)p\}, & (23.116)\\ &\{\gamma p, 1-\gamma p\}, & (23.117) \end{aligned}$$

giving our final expression in the statement of the proposition. □

Exercise 23.6.1 Consider the dephasing channel: $\rho \to (1-p/2)\rho + (p/2)Z\rho Z$. Prove that its quantum capacity is equal to $1 - H_2(p/2)$, where p is the dephasing parameter.

23.7 Discussion of Quantum Capacity

The quantum capacity is particularly well-behaved and understood for the class of degradable channels. Thus, we should not expect any surprises for this class of channels. If a channel is not degradable, we currently cannot say much about the exact value of its quantum capacity, but the study of non-degradable channels has led to many surprises in quantum Shannon theory and this section discusses two of these surprises. The first is the superadditivity of coherent information for the depolarizing channel, and the second is a striking phenomenon known as *superactivation* of quantum capacity, where two channels that individually have zero quantum capacity can combine to make a channel with non-zero quantum capacity.

23.7.1 Superadditivity of Coherent Information

Recall that the depolarizing channel transmits its input with probability $1-p$ and replaces it with the maximally mixed state π with probability p:

$$\rho \to (1-p)\rho + p\pi. \tag{23.118}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We focus on the case where the input and output of this channel is a qubit. The depolarizing channel is an example of a quantum channel that is not degradable.[2] As such, we might expect it to exhibit some strange behavior with respect to its quantum capacity. Indeed, it is known that its coherent information is strictly superadditive when the channel becomes very noisy:

$$5Q(\mathcal{N}) < Q\big(\mathcal{N}^{\otimes 5}\big). \tag{23.119}$$

How can we show that this result is true? First, we can calculate the coherent information of this channel with respect to one channel use. It is possible to show that the maximally entangled state $\Phi^{AA'}$ maximizes the channel coherent information $Q(\mathcal{N})$, and thus

$$Q(\mathcal{N}) = H(B)_{\Phi} - H(AB)_{\mathcal{N}(\Phi)} \tag{23.120}$$

$$= 1 - H(AB)_{\mathcal{N}(\Phi)}, \tag{23.121}$$

where $H(B)_{\Phi} = 1$ follows because the output state on Bob's system is the maximally mixed state whenever the input to the channel is half of a maximally entangled state. In order to calculate $H(AB)_{\mathcal{N}(\Phi)}$, observe that the state on AB is

$$(1-p)\Phi^{AB} + p\pi^{A}\otimes\pi^{B} = (1-p)\Phi^{AB} + \frac{p}{4}I^{AB} \tag{23.122}$$

$$= (1-p)\Phi^{AB} + \frac{p}{4}\big(\big[I^{AB} - \Phi^{AB}\big] + \Phi^{AB}\big) \tag{23.123}$$

$$= \left(1 - \frac{3p}{4}\right)\Phi^{AB} + \frac{p}{4}\big(I^{AB} - \Phi^{AB}\big) \tag{23.124}$$

Since Φ^{AB} and $I^{AB} - \Phi^{AB}$ are orthogonal, the eigenvalues of this state are $1 - 3p/4$ with multiplicity one and $p/4$ with multiplicity three. Thus, the entropy $H(AB)_{\mathcal{N}(\Phi)}$ is

$$H(AB)_{\mathcal{N}(\Phi)} = -\left(1 - \frac{3p}{4}\right)\log\left(1 - \frac{3p}{4}\right) - \frac{3p}{4}\log\Big(\frac{p}{4}\Big), \tag{23.125}$$

and our final expression for the one-shot coherent information is

$$Q(\mathcal{N}) = 1 + \left(1 - \frac{3p}{4}\right)\log\left(1 - \frac{3p}{4}\right) + \frac{3p}{4}\log\Big(\frac{p}{4}\Big). \tag{23.126}$$

Another strategy for transmitting quantum data is to encode half of the maximally entangled state with a five-qubit repetition code:

$$\frac{1}{\sqrt{2}}\Big(|00\rangle^{AA_1} + |11\rangle^{AA_1}\Big)$$

$$\rightarrow \frac{1}{\sqrt{2}}\Big(|000000\rangle^{AA_1A_2A_3A_4A_5} + |111111\rangle^{AA_1A_2A_3A_4A_5}\Big), \tag{23.127}$$

[2] Ref. [232] gives an explicit condition that determines whether a channel is degradable.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

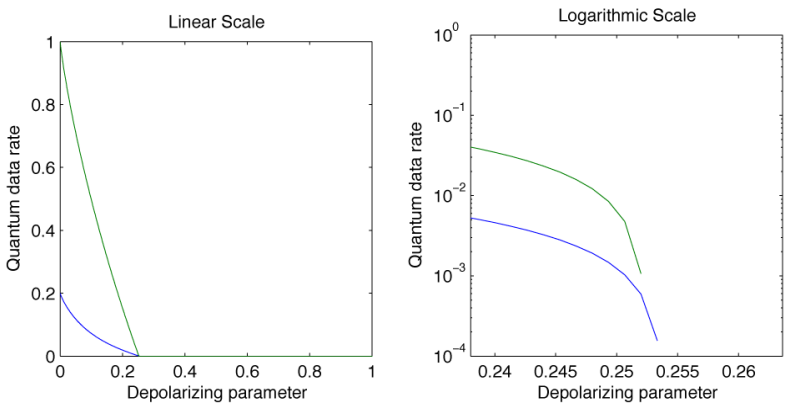

Figure 23.3: The figures plot the coherent information in (23.126) (in green) and that in (23.128) (in blue) versus the depolarizing noise parameter p. The figure on the left is on a linear scale, and the one on the right is on a logarithmic scale. The notable features of the figure on the left are that the quantum data rate of the green curve is equal to one and the quantum data rate of the blue curve is $1/5$ when the channel is noiseless (the latter rate is to be expected for a five-qubit repetition code). Both data rates become small when p is near 0.25, but the figure on the right reveals that the repetition code concatenation strategy still gets positive coherent information even when the rate of the random coding strategy vanishes. This is an example of a channel for which the coherent information can be superadditive.

and calculate the following coherent information with respect to the state resulting from sending the systems $A_1 \cdots A_5$ through the channel:

$$\frac{1}{5} I(A\rangle B_1 B_2 B_3 B_4 B_5). \tag{23.128}$$

(We normalize the above coherent information by five in order to make a fair comparison between a code achieving this rate and one achieving the rate in (23.126).) We know that the rate in (23.128) is achievable by applying the direct part of the quantum capacity theorem to the channel $\mathcal{N}^{\otimes 5}$, and operationally, this strategy amounts to concatenating a random quantum code with a five-qubit repetition code. The remarkable result is that this concatenation strategy can beat the one-shot coherent information when the channel becomes very noisy. Figure 23.3 demonstrates that the concatenation strategy has positive coherent information even when the one-shot coherent information in (23.126) vanishes. This demonstrates superadditivity of coherent information.

Why does this phenomenon occur? The simplest (though perhaps not completely satisfying) explanation is that it results from a phenomenon known as *degeneracy*. Consider a qubit $\alpha|0\rangle + \beta|1\rangle$ encoded in a repetition code:

$$\alpha|00000\rangle + \beta|11111\rangle. \tag{23.129}$$

If the "error" $Z_1 \otimes Z_2$ occurs, then it actually has no effect on this state. The same holds for other two-qubit combinations of Z errors. When the channel noise is low, degeneracy of the code with respect to these errors does not help very much because these two-qubit error

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

combinations are less likely to occur. Though, when the channel becomes really noisy, these errors are more likely to occur, and the help from degeneracy of the repetition code offsets the loss in rate.

It is perhaps strange that the coherent information of a depolarizing channel behaves in this way. The channel seems simple enough, and we could say that the strategies for achieving the unassisted and entanglement-assisted classical capacity of this channel are very "classical" strategies. Recall that the best strategy for achieving the unassisted classical capacity is to generate random codes by picking states uniformly at random from some orthonormal basis, and the receiver measures each channel output in this same orthonormal basis. For achieving the entanglement-assisted classical capacity, we choose a random code by picking Bell states uniformly at random and the receiver measures each channel output and his half of each entangled state in the Bell basis. Both of these results follow from the additivity of the respective capacities. In spite of these other results, the best strategy for achieving the quantum capacity of the depolarizing channel remains very poorly understood.

23.7.2 Superactivation of Quantum Capacity

Perhaps the most startling result in quantum communication is a phenomenon known as *superactivation*. Suppose that Alice is connected to Bob by a quantum channel $\mathcal{N}_1$ with zero capacity for transmitting quantum data. Also, suppose that there is some other zero quantum capacity channel $\mathcal{N}_2$ connecting them. Intuitively, we would expect that Alice should not be able to transmit quantum data reliably over the tensor-product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$. That is, using these channels in parallel seems like it should not give any advantage over using the individual channels alone if they are both individually useless for quantum data transmission (this is the intuition that we have whenever a capacity formula is additive). But two examples of zero-capacity channels are known that can superactivate each other, such that the joint channel has a non-zero quantum capacity. How is this possible?

First, consider a 50% quantum erasure channel $\mathcal{N}_1$ that transmits its input state with probability $1/2$ and replaces it with an erasure state with probability $1/2$. As we have argued before with the no-cloning theorem, such a channel has zero capacity for sending quantum data reliably. Now consider some other channel $\mathcal{N}_2$. We argue in the proof of the following theorem that the coherent information of the joint channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is equal to half the private information of $\mathcal{N}_2$ alone.

Theorem 23.7.1. *Let $\{p_X(x), \rho_x^{A_2}\}$ be an ensemble of inputs for the channel $\mathcal{N}_2$, and let $\mathcal{N}_1$ be a 50% erasure channel. Then there exists a state $\varphi^{A_1A_2}$ such that the coherent information $H(B_1B_2) - H(E_1E_2)$ of the joint channel is equal to half the private information $I(X;B_2) - I(X;E_2)$ of the second channel:*

$$H(B_1B_2)_\omega - H(E_1E_2)_\omega = \frac{1}{2}\Big[I(X;B_2)_\rho - I(X;E_2)_\rho\Big], \tag{23.130}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where

$$\omega^{B_1B_2E_1E_2} \equiv (U_{\mathcal{N}_1} \otimes U_{\mathcal{N}_2})(\varphi^{A_1A_2}), \tag{23.131}$$

$$\rho^{XB_2E_2} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes U_{\mathcal{N}_2}^{A_2\to B_2E_2}(\rho_x^{A_2}), \tag{23.132}$$

and $U_{\mathcal{N}_1}$ and $U_{\mathcal{N}_2}$ are the respective isometric extensions of $\mathcal{N}_1$ and $\mathcal{N}_2$.

Proof. Consider the following classical-quantum state corresponding to the ensemble $\{p_X(x), \rho_x^{A_2}\}$:

$$\rho^{XA_2} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^{A_2}. \tag{23.133}$$

A purification of this state is

$$|\varphi\rangle^{XA_1A_2} \equiv \sum_x \sqrt{p_X(x)}|x\rangle^X(|x\rangle|\phi_x\rangle)^{A_1A_2}, \tag{23.134}$$

where each $|\phi_x\rangle$ is a purification of $\rho_x^{A_2}$ (so that $(|x\rangle|\phi_x\rangle)^{A_1A_2}$ is a purification as well). Let $|\varphi\rangle^{XB_1E_1B_2E_2}$ be the state resulting from sending A_1 and A_2 through the tensor product channel $U_{\mathcal{N}_1} \otimes U_{\mathcal{N}_2}$. We can write this state as follows by recalling the isometric extension of the erasure channel in (23.7):

$$\begin{aligned}|\varphi\rangle^{XB_1E_1B_2E_2} \equiv \frac{1}{\sqrt{2}}\sum_x \sqrt{p_X(x)}|x\rangle^X(|x\rangle|\phi_x\rangle)^{B_1B_2E_2}|e\rangle^{E_1} \\ + \frac{1}{\sqrt{2}}\sum_x \sqrt{p_X(x)}|x\rangle^X(|x\rangle|\phi_x\rangle)^{E_1B_2E_2}|e\rangle^{B_1}.\end{aligned} \tag{23.135}$$

Recall that Bob can perform an isometry on B_1 of the form in (23.89) that identifies whether he receives the state or the erasure symbol, and let Z_B be the classical flag indicating the outcome. Eve can do the same, and let Z_E indicate her flag. Then we can evaluate the coherent information of the state resulting from sending system A_1 through the erasure channel and A_2 through the other channel $\mathcal{N}_2$:

$$\begin{aligned}H(B_1B_2) - H(E_1E_2) &= H(B_1Z_BB_2) - H(E_1Z_EE_2) && (23.136)\\ &= H(B_1B_2|Z_B) + H(Z_B) - H(E_1E_2|Z_E) - H(Z_E) && (23.137)\\ &= H(B_1B_2|Z_B) - H(E_1E_2|Z_E) && (23.138)\\ &= \frac{1}{2}[H(B_2) + H(A_1B_2)] - \frac{1}{2}[H(A_1E_2) + H(E_2)] && (23.139)\\ &= \frac{1}{2}[H(B_2) + H(XE_2)] - \frac{1}{2}[H(XB_2) + H(E_2)] && (23.140)\\ &= \frac{1}{2}[I(X;B_2) - I(X;E_2)]. && (23.141)\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first equality follows because Bob and Eve can perform the isometries that identify whether they receive the state or the erasure flag. The second equality follows from the chaining rule for entropy, and the third follows because the entropies of the flag registers Z_B and Z_E are equal for a 50% erasure channel. The fourth equality follows because the registers Z_B and Z_E are classical, and we can evaluate the conditional entropies as a uniform convex sum of different possibilities: Bob obtaining the state transmitted or not, and Eve obtaining the state transmitted or not. The fifth equality follows because the state on $A_1B_2XE_2$ is pure when conditioning on Bob getting the output of the erasure channel, and the same holds for when Eve gets the output of the erasure channel. The final equality follows from adding and subtracting $H(X)$ and from the definition of quantum mutual information. □

Armed with the above theorem, we need to find an example of a quantum channel that has zero quantum capacity, but for which there exists an ensemble that registers a non-zero private information. If such a channel were to exist, we could combine it with a 50% erasure channel in order to achieve a non-zero coherent information (and thus a non-zero quantum capacity) for the joint channel. Indeed, such a channel exists, and it is known as an entanglement-binding channel. It has the ability to generate private classical communication but no ability to transmit quantum information (we point the reader to Refs. [151, 153] for further details on these channels). Thus, the 50% erasure channel and the entanglement-binding channel can superactivate each other.

The startling phenomenon of superactivation has important implications for quantum data transmission. First, it implies that a quantum channel's ability to transmit quantum information depends on the context in which it is used. For example, if other seemingly useless channels are available, it could be possible to transmit more quantum information than would be possible were the channels used alone. Next, and more importantly for quantum Shannon theory, it implies that whatever formula might eventually be found to characterize quantum capacity (some characterization other than the regularized coherent information in Theorem 23.3.1), it should be strongly non-additive in some cases (strongly non-additive in the sense of superactivation). That is, suppose that $Q^?(\mathcal{N})$ is some unknown formula for the quantum capacity of $\mathcal{N}$ and $Q^?(\mathcal{M})$ is the same formula characterizing the quantum capacity of $\mathcal{M}$. Then this formula in general should be strongly non-additive in some cases:

$$Q^?(\mathcal{N} \otimes \mathcal{M}) > Q^?(\mathcal{N}) + Q^?(\mathcal{M}). \tag{23.142}$$

The discovery of superactivation has led us to realize that at present we are much farther than we might have thought from understanding reliable communication rates over quantum channels.

23.8 Entanglement Distillation

We close out this chaper with a final application of the techniques in the direct coding part of Theorem 23.3.1 to the task of entanglement distillation. Entanglement distillation is a protocol where Alice and Bob begin with many copies of some bipartite state ρ^{AB}. They

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

attempt to distill ebits from it at some positive rate by employing local operations and forward classical communication from Alice to Bob. If the state is pure, then Alice and Bob should simply perform the entanglement concentration protocol from Chapter 18, and there is no need for forward classical communication in this case. Otherwise, they can perform the protocol given in the proof of the following theorem.

Theorem 23.8.1 (Devetak-Winter)**.** *Suppose that Alice and Bob share the state $(\rho^{AB})^{\otimes n}$ where n is an arbitrarily large number. Then it is possible for them to distill ebits at the rate $I(A\rangle B)_\rho$ if they are allowed forward classical communication from Alice to Bob.*

We should mention that we have already proved the statement in the above theorem with the protocol given in Corollary 21.4.2. Nevertheless, it is still instructive to exploit the techniques from this chapter in proving the existence of an entanglement distillation protocol.

Proof. Suppose that Alice and Bob begin with a general bipartite state ρ^{AB} with purification ψ^{ABE}. We can write the purification in Schmidt form as follows:

$$|\psi\rangle^{ABE} \equiv \sum_{x\in\mathcal{X}} \sqrt{p_X(x)}|x\rangle^A \otimes |\psi_x\rangle^{BE}. \tag{23.143}$$

The n^{th} extension of the above state is

$$|\psi\rangle^{A^nB^nE^n} \equiv \sum_{x^n\in\mathcal{X}^n} \sqrt{p_{X^n}(x^n)}|x^n\rangle^{A^n} \otimes |\psi_{x^n}\rangle^{B^nE^n}. \tag{23.144}$$

The protocol begins with Alice performing a type class measurement given by the type projectors (recall from (14.118) that the typical projector decomposes into a sum of the type class projectors):

$$\Pi_t^n \equiv \sum_{x^n\in T_t^{X^n}} |x^n\rangle\langle x^n|. \tag{23.145}$$

If the type resulting from the measurement is not a typical type, then Alice aborts the protocol (this result happens with arbitrarily small probability). If it is a typical type, they can then consider a code over a particular type class t with the following structure:

$$LMK \approx |T_t| \approx 2^{nH(X)}, \tag{23.146}$$

$$K \approx 2^{nI(X;E)}, \tag{23.147}$$

$$MK \approx 2^{nI(X;B)}, \tag{23.148}$$

where t is the type class and the entropies are with respect to the following dephased state:

$$\sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x|^X \otimes |\psi_x\rangle\langle\psi_x|^{BE}. \tag{23.149}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

It follows that $M \approx 2^{n(I(X;B)-I(X;E))} = 2^{n[H(B)-H(E)]}$ and $L \approx 2^{nH(X|B)}$. We label the codewords as $x^n(l,m,k)$ where $x^n(l,m,k) \in T_t$. Thus, we instead operate on the following state $|\tilde{\psi}_t\rangle^{A^nB^nE^n}$ resulting from the type class measurement:

$$|\tilde{\psi}_t\rangle^{A^nB^nE^n} \equiv \frac{1}{\sqrt{|T_t|}} \sum_{x^n \in T_t} |x^n\rangle^{A^n} \otimes |\psi_{x^n}\rangle^{B^nE^n}. \tag{23.150}$$

The protocol proceeds as follows. Alice first performs the following incomplete measurement of the system A^n:

$$\left\{ \Gamma_l \equiv \sum_{m,k} |m,k\rangle\langle x^n(l,m,k)|^{A^n} \right\}_l. \tag{23.151}$$

This measurement collapses the above state as follows:

$$\frac{1}{\sqrt{MK}} \sum_{m,k} |m,k\rangle^{A^n} \otimes \left|\psi_{x^n(l,m,k)}\right\rangle^{B^nE^n}. \tag{23.152}$$

Alice transmits the classical information in l to Bob, using $nH(X|B)$ bits of classical information. Bob needs to know l so that he can know in which code they are operating. Bob then constructs the following isometry, a coherent POVM similar to that in (23.30) (constructed from the POVM for a private classical communication code):

$$\sum_{m,k} \sqrt{\Lambda}^{B^n}_{m,k} \otimes |m,k\rangle^B. \tag{23.153}$$

After performing the above coherent POVM, his state is close to the following one:

$$\frac{1}{\sqrt{MK}} \sum_{m,k} |m,k\rangle^{A^n} \otimes |m,k\rangle^B \left|\psi_{x^n(l,m,k)}\right\rangle^{B^nE^n}. \tag{23.154}$$

Alice then performs a measurement of the k register in the Fourier-transformed basis:

$$\left\{ |\hat{t}\rangle \equiv \frac{1}{\sqrt{K}} \sum_k e^{i2\pi kt/K} |k\rangle \right\}_{t \in \{1,\ldots,K\}}. \tag{23.155}$$

Alice performs this particular measurement because she would like Bob and Eve to maintain their entanglement in the k variable. The state resulting from this measurement is

$$\frac{1}{\sqrt{MK}} \sum_{m,k} |m\rangle^{A^n} \otimes e^{i2\pi kt/K} |m,k\rangle^B \left|\psi_{x^n(l,m,k)}\right\rangle^{B^nE^n}. \tag{23.156}$$

Alice then uses $nI(X;E)$ bits to communicate the t variable to Bob. Bob then applies the phase transformation $Z^\dagger(t)$, where

$$Z^\dagger(t) = \sum_k e^{-i2\pi tk/K} |k\rangle\langle k|, \tag{23.157}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

to his k variable in register B. The resulting state is

$$\frac{1}{\sqrt{MK}} \sum_{m,k} |m\rangle^{A^n} \otimes |m,k\rangle^{B} \left|\psi_{x^n(l,m,k)}\right\rangle^{B^n E^n}. \tag{23.158}$$

They then proceed as in the final steps (23.47-23.54) of the protocol from the direct coding part of Theorem 23.3.1, and they extract a state close to a maximally entangled state of the following form:

$$\frac{1}{\sqrt{M}} \sum_{m} |m\rangle^{A^n} \otimes |m\rangle^{B}, \tag{23.159}$$

with rate equal to $(\log M)/n = H(B) - H(E)$. □

Exercise 23.8.1 Argue that the above protocol cannot perform the task of state transfer as can the protocol in Corollary 21.4.2.

23.9 History and Further Reading

The quantum capacity theorem has a long history that led to many important discoveries in quantum information theory. Shor first stated the problem of finding the quantum capacity of a quantum channel in his seminal paper on quantum error correction [224]. DiVincenzo *et al.* demonstrated that the coherent information of the depolarizing channel is superadditive by concatenating a random code with a repetition code [80] (this result in hindsight was remarkable given that the coherent information was not even known at the time). Smith and Smolin later extended this result to show that the coherent information is strongly superadditive for several examples of Pauli channels [232]. Schumacher and Nielsen demonstrated that the coherent information obeys a quantum data processing inequality [218], much like the classical data processing inequality for mutual information. Schumacher and Westmoreland started making connections between quantum privacy and quantum coherence [220]. Bennett *et al.* [30] and Barnum *et al.* [15] demonstrated that forward classical communication cannot increase the quantum capacity. In the same paper, Bennett *et al.* [30] introduced the idea of entanglement distillation, which has important connections with the quantum capacity.

Barnum, Knill, Nielsen, and Schumacher made important progress on the quantum capacity theorem in a series of papers that established the coherent information upper bound on the quantum capacity [217, 218, 16, 15]. Lloyd [185], Shor [227], and Devetak [68] are generally credited with proving the coherent information lower bound on the quantum capacity, though an inspection of Lloyd's proof reveals that it is perhaps not as rigorous as the latter two proofs. Shor delivered his proof of the lower bound in a lecture [227], though he never published this proof in a journal. Later, Hayden, Shor, and Winter published a paper [134] detailing a proof of the quantum capacity theorem that they considered to be closest in spirit to Shor's proof in Ref. [227]. After Shor's proof, Devetak provided a fully rigorous proof of the lower bound on the quantum capacity [68], by analyzing superpositions of the codewords

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

from private classical codes. This is the approach we have taken in this chapter. We should also mention that Hamada showed how to achieve the coherent information for certain input states by using random stabilizer codes [121], and Harrington and Preskill showed how to achieve the coherent information rate for a very specific class of channels [122].

Another approach to proving the quantum capacity theorem is known as the decoupling approach [132]. This approach exploits a fundamental concept introduced by Schumacher and Westmoreland in Ref. [221]. Suppose that the reference, Bob, and Eve share a tripartite pure entangled state $|\psi\rangle^{RBE}$ after Alice transmits her share of the entanglement with the reference through a noisy channel. Then if the reduced state ψ^{RE} on the reference system and Eve's system is approximately decoupled, meaning that

$$\left\|\psi^{RE} - \psi^{R} \otimes \sigma^{E}\right\|_1 \leq \epsilon, \tag{23.160}$$

where σ^E is some arbitrary state, this implies that Bob can decode the quantum information that Alice intended to send to him. Why is this so? Let's suppose that the state is exactly decoupled. Then one purification of the state ψ^{RE} is the state $|\psi\rangle^{RBE}$ that they share after the channel acts. Another purification of $\psi^{RE} = \psi^R \otimes \sigma^E$ is

$$|\psi\rangle^{RB_1} \otimes |\sigma\rangle^{B_2E}, \tag{23.161}$$

where $|\psi\rangle^{RB_1}$ is the original state that Alice sent through the channel and $|\sigma\rangle^{B_2E}$ is some other state that purifies the state σ^E of the environment. Since all purifications are related by isometries and since Bob possesses the purification of R and E, there exists some unitary $U^{B\to B_1B_2}$ such that

$$U^{B\to B_1B_2}|\psi\rangle^{RBE} = |\psi\rangle^{RB_1} \otimes |\sigma\rangle^{B_2E}. \tag{23.162}$$

This unitary is then Bob's decoder! Thus, the decoupling condition implies the existence of a decoder for Bob, so that it is only necessary to show the existence of an encoder that decouples the reference from the environment. Simply put, the structure of quantum mechanics allows for this way of proving the quantum capacity theorem.

Many researchers have now exploited the decoupling approach in a variety of contexts. This approach is implicit in Devetak's proof of the quantum capacity theorem [68]. Horodecki *et al.* exploited it to prove the existence of a state merging protocol [148, 149]. Yard and Devetak [267] and Ye *et al.* [271] used it in their proofs of the state redistribution protocol. Dupuis *et al.* [84] proved the best known characterization of the entanglement-assisted quantum capacity of the broadcast channel with this approach. The thesis of Dupuis and subsequent work generalize this decoupling approach to settings beyond the traditional IID setting [82, 83]. Datta and coworkers have also applied this approach in a variety of contexts [50, 63, 62], and Ref. [251] used the approach to study quantum communication with a noisy channel and a noisy state.

Bennett *et al.* found the quantum capacity of the erasure channel in Ref. [29], and Fazio and Giovannetti computed the quantum capacity of the amplitude damping channel in Ref. [102]. Smith *et al.* showed superactivation in Ref. [234] and later showed superactivation for channels that can be realized more easily in the laboratory [233]. Devetak and Winter established that the coherent information is achievable for entanglement distillation [76].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 24

Trading Resources for Communication

This chapter unifies all of the channel coding theorems that we have studied in this book. One of the most general information processing tasks that a sender and receiver can accomplish is to transmit classical and quantum information and generate entanglement with many independent uses of a quantum channel and with the assistance of classical communication, quantum communication, and shared entanglement.[1] The resulting rates for communication are *net* rates that give the generation rate of a resource less its consumption rate. Since we have three resources, all achievable rates are rate triples (C, Q, E) that lie in a three-dimensional capacity region, where C is the net rate of classical communication, Q is the net rate of quantum communication, and E is the net rate of entanglement consumption/generation. The capacity theorem for this general scenario is known as the quantum dynamic capacity theorem, and it is the main theorem that we prove in this chapter. All of the rates given in the channel coding theorems of previous chapters are special points in this three-dimensional capacity region.

The proof of the quantum dynamic capacity theorem comes in two parts: the direct coding theorem and the converse theorem. The direct coding theorem demonstrates that the strategy for achieving any point in the three-dimensional capacity region is remarkably simple: we just combine the protocol from Corollary 21.5.2 for entanglement-assisted classical and quantum communication with the three unit protocols of teleportation, super-dense coding, and entanglement distribution. The interpretation of the achievable rate region is that it is the unit resource capacity region from Chapter 8 translated along the points achievable with the protocol from Corollary 21.5.2. The proof of the converse theorem is perhaps the more difficult part—we analyze the most general protocol that can consume and generate classical communication, quantum communication, and entanglement along with the consumption of many independent uses of a quantum channel, and we show that the net rates for such a protocol are bounded by the achievable rate region. In the general case, our characterization is multi-letter, meaning that the computation of the capacity region requires

[1] Recall that Chapter 8 addressed a special case of this information processing task that applies to the scenario in which the sender and receiver do not have access to many independent uses of a noisy quantum channel.

an optimization over a potentially infinite number of channel uses and is thus intractable. Though, the quantum Hadamard channels from Section 5.2.4 are a special class of channels for which the regularization is not necessary, and we can compute their capacity regions over a single instance of the channel. Another important class of channels for which the capacity region is known is the class of lossy bosonic channels (though the optimality proof is only up to a long-standing conjecture which many researchers believe to be true). These lossy bosonic channels model free-space communication or loss in a fiber optic cable and thus have an elevated impetus for study because of their importance in practical applications.

One of the most important questions for communication in this three-dimensional setting is whether it is really necessary to exploit the trade-off coding strategy given in Corollary 21.5.2. That is, would it be best simply to use a classical communication code for a fraction of the channel uses, a quantum communication code for another fraction, an entanglement-assisted code for another fraction, etc.? Such a strategy is known as time-sharing and allows the sender and receiver to achieve convex combinations of any rate triples in the capacity region. The answer to this question depends on the channel. For example, time-sharing is optimal for the quantum erasure channel, but it is not for a dephasing channel or a lossy bosonic channel. In fact, trade-off coding for a lossy bosonic channel can give tremendous performance gains over time-sharing. How can we know which one will perform better in the general case? It is hard to say, but at the very least, we know that time-sharing is a special case of trade-off coding as we argued in Section 21.5.2. Thus, from this perspective, it might make sense simply to always use a trade-off strategy.

We organize this chapter as follows. We first review the information processing task corresponding to the quantum dynamic capacity region. Section 24.2 states the quantum dynamic capacity theorem and shows how many of the capacity theorems we studied previously arise as special cases of it. The next two sections prove the direct coding theorem and the converse theorem. Section 24.4.2 introduces the quantum dynamic capacity formula, which is important for analyzing whether the quantum dynamic capacity region is single-letter. In the final section of this chapter, we compute and plot the quantum dynamic capacity region for the dephasing channels and the lossy bosonic channels.

24.1 The Information Processing Task

Figure 24.1 depicts the most general protocol for generating classical communication, quantum communication, and entanglement with the consumption of a noisy quantum channel $\mathcal{N}^{A'\to B}$ and the same respective resources. Alice possesses two classical registers (each labeled by M and of dimension $2^{n\bar{C}}$), a quantum register A_1 of dimension $2^{n\bar{Q}}$ entangled with a reference system R, and another quantum register T_A of dimension $2^{n\tilde{E}}$ that contains her half of the shared entanglement with Bob:

$$\omega^{MMRA_1T_AT_B} \equiv \overline{\Phi}^{MM} \otimes \Phi^{RA_1} \otimes \Phi^{T_AT_B}. \tag{24.1}$$

She passes one of the classical registers and the registers A_1 and T_A into a CPTP encoding map $\mathcal{E}^{MA_1T_A\to A'^nS_ALA_2}$ that outputs a quantum register S_A of dimension $2^{n\bar{E}}$ and a quantum

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

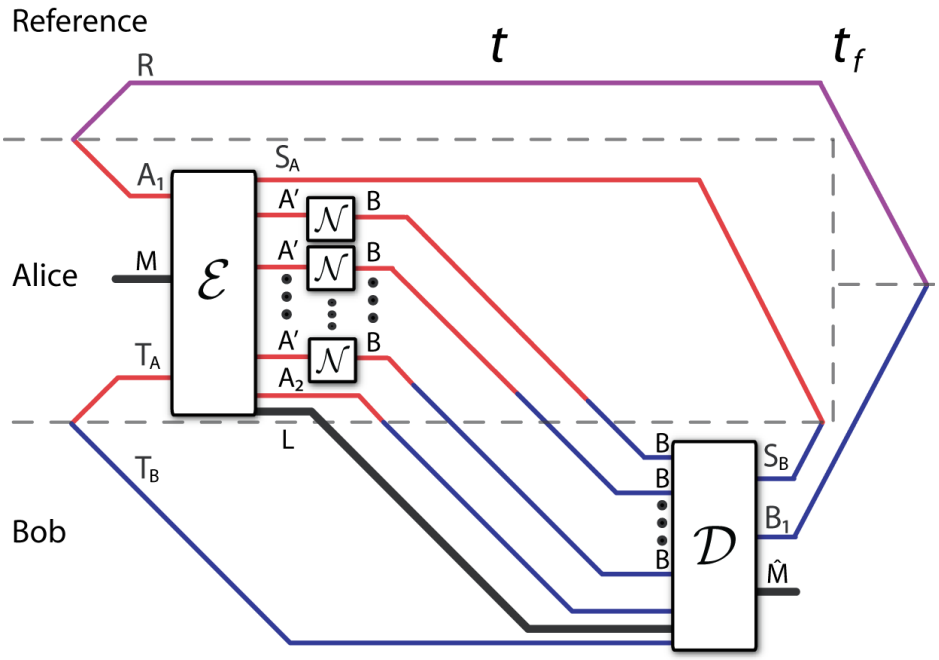

Figure 24.1: The most general protocol for generating classical communication, quantum communication, and entanglement generation with the help of the same respective resources and many uses of a noisy quantum channel. Alice begins with her classical register M, her quantum register A_1, and her half of the shared entanglement in register T_A. She encodes according to some CPTP map $\mathcal{E}$ that outputs a quantum register S_A, many registers A'^n, a quantum register A_2, and a classical register L. She inputs A'^n to many uses of the noisy channel $\mathcal{N}$ and transmits A_2 over a noiseless quantum channel and L over a noiseless classical channel. Bob receives the channel outputs B^n, the quantum register A_2, and the classical register L and performs a decoding $\mathcal{D}$ that recovers the quantum information and classical message. The decoding also generates entanglement with system S_A. Many protocols are a special case of the above one. For example, the protocol is entanglement-assisted communication of classical and quantum information if the registers L, S_A, S_B, and A_2 are null.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

register A_2 of dimension $2^{n\tilde{Q}}$, a classical register L of dimension $2^{n\tilde{C}}$, and many quantum systems A'^n for input to the channel. The register S_A is for creating entanglement with Bob. The state after the encoding map $\mathcal{E}$ is as follows:

$$\omega^{MA'^nS_ALA_2RT_B} \equiv \mathcal{E}^{MA_1T_A\to A'^nS_ALA_2}(\omega^{MMRA_1T_AT_B}). \tag{24.2}$$

She sends the systems A'^n through many uses $\mathcal{N}^{A'^n\to B^n}$ of the noisy channel $\mathcal{N}^{A'\to B}$, transmits L over a noiseless classical channel, and transmits A_2 over a noiseless quantum channel, producing the following state:

$$\omega^{MB^nS_ALA_2RT_B} \equiv \mathcal{N}^{A'^n\to B^n}(\omega^{MA'^nS_ALA_2RT_B}). \tag{24.3}$$

The above state is a state of the following form:

$$\sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A'^n\to B^n}(\rho_x^{AA'^n}), \tag{24.4}$$

with $A \equiv RT_BA_2S_A$ and $X \equiv ML$. Bob then applies a map $\mathcal{D}^{B^nA_2T_BL\to B_1S_B\hat{M}}$ that outputs a quantum system B_1, a quantum system S_B, and a classical register $\hat{M}$. Let ω' denote the final state. The following condition holds for a good protocol:

$$\left\| \overline{\Phi}^{M\hat{M}} \otimes \Phi^{RB_1} \otimes \Phi^{S_AS_B} - (\omega')^{MB_1S_B\hat{M}S_AR} \right\|_1 \leq \epsilon, \tag{24.5}$$

implying that Alice and Bob establish maximal classical correlations in M and $\hat{M}$ and maximal entanglement between S_A and S_B. The above condition also implies that the coding scheme preserves the entanglement with the reference system R. The net rate triple for the protocol is as follows: $\left(\bar{C} - \tilde{C} - \delta, \bar{Q} - \tilde{Q} - \delta, \bar{E} - \tilde{E} - \delta\right)$ for some arbitrarily small $\delta > 0$. The protocol generates a resource if its corresponding rate is positive, and it consumes a resource if its corresponding rate is negative. We say that a rate triple (C, Q, E) is achievable if there exists a protocol of the above form for all $\delta, \epsilon > 0$ and sufficiently large n.

24.2 The Quantum Dynamic Capacity Theorem

The dynamic capacity theorem gives bounds on the reliable communication rates of a noisy quantum channel when combined with the noiseless resources of classical communication, quantum communication, and shared entanglement. The theorem applies regardless of whether a protocol consumes the noiseless resources or generates them.

Theorem 24.2.1 (Quantum Dynamic Capacity)**.** *The dynamic capacity region $\mathcal{C}_{\text{CQE}}(\mathcal{N})$ of a quantum channel $\mathcal{N}$ is equal to the following expression:*

$$\mathcal{C}_{\text{CQE}}(\mathcal{N}) = \overline{\bigcup_{k=1}^{\infty} \frac{1}{k}\mathcal{C}^{(1)}_{\text{CQE}}(\mathcal{N}^{\otimes k})}, \tag{24.6}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where the overbar indicates the closure of a set. The "one-shot" region $\mathcal{C}^{(1)}_{\mathrm{CQE}}(\mathcal{N})$ *is the union of the "one-shot, one-state" regions* $\mathcal{C}^{(1)}_{\mathrm{CQE},\sigma}(\mathcal{N})$*:*

$$\mathcal{C}^{(1)}_{\mathrm{CQE}}(\mathcal{N}) \equiv \bigcup_{\sigma} \mathcal{C}^{(1)}_{\mathrm{CQE},\sigma}(\mathcal{N}). \tag{24.7}$$

The "one-shot, one-state" region $\mathcal{C}^{(1)}_{\mathrm{CQE},\sigma}(\mathcal{N})$ *is the set of all rates* C*,* Q*, and* E*, such that*

$$C + 2Q \leq I(AX;B)_\sigma, \tag{24.8}$$
$$Q + E \leq I(A\rangle BX)_\sigma, \tag{24.9}$$
$$C + Q + E \leq I(X;B)_\sigma + I(A\rangle BX)_\sigma. \tag{24.10}$$

The above entropic quantities are with respect to a classical-quantum state σ^{XAB} *where*

$$\sigma^{XAB} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A'\to B}(\phi_x^{AA'}), \tag{24.11}$$

and the states $\phi_x^{AA'}$ *are pure. It is implicit that one should consider states on* A'^k *instead of* A' *when taking the regularization in (24.6).*

The above theorem is a "multi-letter" capacity theorem because of the regularization in (24.6). Though, we show in Section 24.5.1 that the regularization is not necessary for the Hadamard class of channels. We prove the above theorem in two parts:

1. The direct coding theorem in Section 24.3 shows that combining the protocol from Corollary 21.5.2 with teleportation, super-dense coding, and entanglement distribution achieves the above region.

2. The converse theorem in Section 24.4 demonstrates that any coding scheme cannot do better than the regularization in (24.6), in the sense that a scheme with vanishing error should have its rates below the above amounts.

Exercise 24.2.1 Show that it suffices to evaluate just the following four entropies in order to determine the one-shot, one-state region in Theorem 24.2.1:

$$H(A|X)_\sigma = \sum_x p_X(x) H(A)_{\phi_x}, \tag{24.12}$$
$$H(B)_\sigma = H\left(\sum_x p_X(x)\mathcal{N}^{A'\to B}(\phi_x^{A'})\right), \tag{24.13}$$
$$H(B|X)_\sigma = \sum_x p_X(x) H\left(\mathcal{N}^{A'\to B}(\phi_x^{A'})\right), \tag{24.14}$$
$$H(E|X)_\sigma = \sum_x p_X(x) H\left((\mathcal{N}^c)^{A'\to E}(\phi_x^{A'})\right), \tag{24.15}$$

where the state σ^{XAB} is of the form in (24.11).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

24.2.1 Special Cases of the Quantum Dynamic Capacity Theorem

We first consider five special cases of the above capacity theorem that arise when Q and E both vanish, C and E both vanish, or one of C, Q, or E vanishes. The first two cases correspond respectively to the classical capacity theorem from Chapter 19 and the quantum capacity theorem from Chapter 23. Each of the other special cases traces out a two-dimensional achievable rate region in the three-dimensional capacity region. The five coding scenarios are as follows:

1. Classical communication (C) when there is no entanglement assistance or quantum communication. The achievable rate region lies on the $(C, 0, 0)$ ray extending from the origin.

2. Quantum communication (Q) when there is no entanglement assistance or classical communication. The achievable rate region lies on the $(0, Q, 0)$ ray extending from the origin.

3. Entanglement-assisted quantum communication (QE) when there is no classical communication. The achievable rate region lies in the $(0, Q, -E)$ quarter-plane of the three-dimensional region in Theorem 24.2.1.

4. Classically-enhanced quantum communication (CQ) when there is no entanglement assistance. The achievable rate region lies in the $(C, Q, 0)$ quarter-plane of the three-dimensional region in Theorem 24.2.1.

5. Entanglement-assisted classical communication (CE) when there is no quantum communication. The achievable rate region lies in the $(C, 0, -E)$ quarter-plane of the three-dimensional region in Theorem 24.2.1.

Classical Capacity

The following theorem gives the one-dimensional capacity region $\mathcal{C}_{\mathrm{C}}(\mathcal{N})$ of a quantum channel $\mathcal{N}$ for classical communication.

Theorem 24.2.2 (Holevo-Schumacher-Westmoreland)**.** *The classical capacity region $\mathcal{C}_{\mathrm{C}}(\mathcal{N})$ is given by*

$$\mathcal{C}_{\mathrm{C}}(\mathcal{N}) = \overline{\bigcup_{k=1}^{\infty} \frac{1}{k} \mathcal{C}_{\mathrm{C}}^{(1)}(\mathcal{N}^{\otimes k})}. \tag{24.16}$$

The "one-shot" region $\mathcal{C}_{\mathrm{C}}^{(1)}(\mathcal{N})$ is the union of the regions $\mathcal{C}_{\mathrm{C},\sigma}^{(1)}(\mathcal{N})$, where $\mathcal{C}_{\mathrm{C},\sigma}^{(1)}(\mathcal{N})$ is the set of all $C \geq 0$, such that

$$C \leq I(X;B)_{\sigma} + I(A\rangle BX)_{\sigma}. \tag{24.17}$$

The entropic quantity is with respect to the state σ^{XABE} in (24.11).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The bound in (24.17) is a special case of the bound in (24.10) with $Q = 0$ and $E = 0$. The above characterization of the classical capacity region may seem slightly different from the characterization in Chapter 19, until we make a few observations. First, we rewrite the coherent information $I(A\rangle BX)_\sigma$ as $H(B|X)_\sigma - H(E|X)_\sigma$. Then $I(X;B)_\sigma + I(A\rangle BX)_\sigma = H(B)_\sigma - H(E|X)_\sigma$. Next, pure states of the form $|\varphi\rangle_x^{A'}$ are sufficient to attain the classical capacity of a quantum channel (see Theorem 12.3.2). We briefly recall this argument. An ensemble of the following form realizes the classical capacity of a quantum channel:

$$\rho^{XA'} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^{A'}. \tag{24.18}$$

This ensemble itself is a restriction of the ensemble in (24.11) to the systems X and A'. Each mixed state $\rho_x^{A'}$ admits a spectral decomposition of the form $\rho_x^{A'} = \sum_y p_{Y|X}(y|x)\psi_{x,y}^{A'}$ where $\psi_{x,y}^{A'}$ is a pure state. We can define an augmented classical-quantum state $\theta^{XYA'}$ as follows:

$$\theta^{XYA'} \equiv \sum_{x,y} p_{Y|X}(y|x)p_X(x)|x\rangle\langle x|^X \otimes |y\rangle\langle y|^Y \otimes \psi_{x,y}^{A'}, \tag{24.19}$$

so that $\mathrm{Tr}_Y\{\theta^{XYA'}\} = \rho^{XA'}$. Sending the A' system of the states $\rho^{XA'}$ and $\theta^{XYA'}$ leads to the respective states ρ^{XB} and θ^{XYB}. Then the following equality and inequality hold

$$I(X;B)_\rho = I(X;B)_\theta \tag{24.20}$$

$$\leq I(XY;B)_\theta, \tag{24.21}$$

where the equality holds because $\mathrm{Tr}_Y\{\theta^{XYA'}\} = \rho^{XA'}$ and the inequality follows from quantum data processing. Redefining the classical variable as the joint random variable X, Y reveals that it is sufficient to consider pure state ensembles for the classical capacity. Returning to our main argument, then $H(E|X)_\sigma = H(B|X)_\sigma$ so that $I(X;B)_\sigma + I(A\rangle BX)_\sigma = H(B)_\sigma - H(B|X)_\sigma = I(X;B)_\sigma$ for states of this form. Thus, the expression in (24.17) can never exceed the classical capacity and finds its maximum exactly at the Holevo information.

Quantum Capacity

The following theorem gives the one-dimensional quantum capacity region $\mathcal{C}_{\mathrm{Q}}(\mathcal{N})$ of a quantum channel $\mathcal{N}$.

Theorem 24.2.3 (Quantum Capacity)**.** *The quantum capacity region $\mathcal{C}_{\mathrm{Q}}(\mathcal{N})$ is given by*

$$\mathcal{C}_{\mathrm{Q}}(\mathcal{N}) = \overline{\bigcup_{k=1}^{\infty} \frac{1}{k}\mathcal{C}_{\mathrm{Q}}^{(1)}(\mathcal{N}^{\otimes k})}. \tag{24.22}$$

The "one-shot" region $\mathcal{C}_{\mathrm{Q}}^{(1)}(\mathcal{N})$ is the union of the regions $\mathcal{C}_{\mathrm{Q},\sigma}^{(1)}(\mathcal{N})$, where $\mathcal{C}_{\mathrm{Q},\sigma}^{(1)}(\mathcal{N})$ is the set of all $Q \geq 0$, such that

$$Q \leq I(A\rangle BX)_\sigma. \tag{24.23}$$

The entropic quantity is with respect to the state σ^{XABE} in (24.11) with the restriction that the density $p_X(x)$ is degenerate.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The bound in (24.23) is a special case of the bound in (24.9) with $E = 0$. The other bounds in Theorem 24.2.1 are looser than the bound in (24.9) when $C, E = 0$.

Entanglement-Assisted Quantum Capacity

The following theorem gives the two-dimensional entanglement-assisted quantum capacity region $\mathcal{C}_{\text{QE}}(\mathcal{N})$ of a quantum channel $\mathcal{N}$.

Theorem 24.2.4 (Devetak-Harrow-Winter)**.** *The entanglement-assisted quantum capacity region $\mathcal{C}_{\text{QE}}(\mathcal{N})$ is given by*

$$\mathcal{C}_{\text{QE}}(\mathcal{N}) = \overline{\bigcup_{k=1}^{\infty} \frac{1}{k} \mathcal{C}^{(1)}_{\text{QE}}(\mathcal{N}^{\otimes k})}. \tag{24.24}$$

The "one-shot" region $\mathcal{C}^{(1)}_{\text{QE}}(\mathcal{N})$ is the union of the regions $\mathcal{C}^{(1)}_{\text{QE},\sigma}(\mathcal{N})$, where $\mathcal{C}^{(1)}_{\text{QE},\sigma}(\mathcal{N})$ is the set of all $Q, E \geq 0$, such that

$$2Q \leq I(AX;B)_\sigma, \tag{24.25}$$

$$Q \leq I(A\rangle BX)_\sigma + |E|. \tag{24.26}$$

The entropic quantities are with respect to the state σ^{XABE} in (24.11) with the restriction that the density $p_X(x)$ is degenerate.

The bounds in (24.25) and (24.26) are a special case of the respective bounds in (24.8) and (24.9) with $C = 0$. The other bounds in Theorem 24.2.1 are looser than the bounds in (24.8) and (24.9) when $C = 0$. Observe that the region is a union of general pentagons (see the QE-plane in Figure 24.2 for an example of one of these general pentagons in the union).

Classically-Enhanced Quantum Capacity

The following theorem gives the two-dimensional capacity region $\mathcal{C}_{CQ}(\mathcal{N})$ for classically-enhanced quantum communication over a quantum channel $\mathcal{N}$.

Theorem 24.2.5 (Devetak-Shor)**.** *The classically-enhanced quantum capacity region $\mathcal{C}_{\text{CQ}}(\mathcal{N})$ is given by*

$$\mathcal{C}_{\text{CQ}}(\mathcal{N}) = \overline{\bigcup_{k=1}^{\infty} \frac{1}{k} \mathcal{C}^{(1)}_{\text{CQ}}(\mathcal{N}^{\otimes k})}. \tag{24.27}$$

The "one-shot" region $\mathcal{C}^{(1)}_{\text{CQ}}(\mathcal{N})$ is the union of the regions $\mathcal{C}^{(1)}_{\text{CQ},\sigma}(\mathcal{N})$, where $\mathcal{C}^{(1)}_{\text{CQ},\sigma}(\mathcal{N})$ is the set of all $C, Q \geq 0$, such that

$$C + Q \leq I(X;B)_\sigma + I(A\rangle BX)_\sigma, \tag{24.28}$$

$$Q \leq I(A\rangle BX)_\sigma. \tag{24.29}$$

The entropic quantities are with respect to the state σ^{XABE} in (24.11).

The bounds in (24.28) and (24.29) are a special case of the respective bounds in (24.9) and (24.10) with $E = 0$. Observe that the region is a union of trapezoids (see the CQ-plane in Figure 24.2 for an example of one of these rectangles in the union).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Entanglement-Assisted Classical Capacity with Limited Entanglement

Theorem 24.2.6 (Shor)**.** *The entanglement-assisted classical capacity region $\mathcal{C}_{\text{CE}}(\mathcal{N})$ of a quantum channel $\mathcal{N}$ is*

$$\mathcal{C}_{\text{CE}}(\mathcal{N}) = \overline{\bigcup_{k=1}^{\infty} \frac{1}{k} \mathcal{C}_{\text{CE}}^{(1)}(\mathcal{N}^{\otimes k})}. \tag{24.30}$$

The "one-shot" region $\mathcal{C}_{\text{CE}}^{(1)}(\mathcal{N})$ is the union of the regions $\mathcal{C}_{\text{CE},\sigma}^{(1)}(\mathcal{N})$, where $\mathcal{C}_{\text{CE},\sigma}^{(1)}(\mathcal{N})$ is the set of all $C, E \geq 0$, such that

$$C \leq I(AX;B)_\sigma, \tag{24.31}$$
$$C \leq I(X;B)_\sigma + I(A\rangle BX)_\sigma + |E|, \tag{24.32}$$

where the entropic quantities are with respect to the state σ^{XABE} in (24.11).

The bounds in (24.31) and (24.32) are a special case of the respective bounds in (24.8) and (24.10) with $Q = 0$. Observe that the region is a union of general polyhedra (see the CE-plane in Figure 24.2 for an example of one of these general polyhedra in the union).

24.3 The Direct Coding Theorem

The unit resource achievable region is what Alice and Bob can achieve with the protocols entanglement distribution, teleportation, and super-dense coding (see Chapter 8). It is the cone of the rate triples corresponding to these protocols:

$$\{\alpha(0,-1,1) + \beta(2,-1,-1) + \gamma(-2,1,-1) : \alpha, \beta, \gamma \geq 0\}. \tag{24.33}$$

We can also write any rate triple (C, Q, E) in the unit resource capacity region with a matrix equation:

$$\begin{bmatrix} C \\ Q \\ E \end{bmatrix} = \begin{bmatrix} 0 & 2 & -2 \\ -1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}. \tag{24.34}$$

The inverse of the above matrix is as follows:

$$\begin{bmatrix} -\frac{1}{2} & -1 & 0 \\ 0 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}, \tag{24.35}$$

and gives the following set of inequalities for the unit resource achievable region:

$$C + 2Q \leq 0, \tag{24.36}$$
$$Q + E \leq 0, \tag{24.37}$$
$$C + Q + E \leq 0, \tag{24.38}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

by inverting the matrix equation in (24.34) and applying the constraints $\alpha, \beta, \gamma \geq 0$.

Now, let us include the protocol from Corollary 21.5.2 for entanglement-assisted communication of classical and quantum information. Corollary 21.5.2 states that we can achieve the following rate triple by channel coding over a noisy quantum channel $\mathcal{N}^{A'\to B}$:

$$\left(I(X;B)_\sigma, \frac{1}{2} I(A;B|X)_\sigma, -\frac{1}{2} I(A;E|X)_\sigma \right), \tag{24.39}$$

for any state σ^{XABE} of the form:

$$\sigma^{XABE} \equiv \sum_x p_X(x) |x\rangle\langle x|^X \otimes U_{\mathcal{N}}^{A'\to BE}(\phi_x^{AA'}), \tag{24.40}$$

where $U_{\mathcal{N}}^{A'\to BE}$ is an isometric extension of the quantum channel $\mathcal{N}^{A'\to B}$. Specifically, we showed in Corollary 21.5.2 that one can achieve the above rates with vanishing error in the limit of large blocklength. Thus the achievable rate region is the following translation of the unit resource achievable region in (24.34):

$$\begin{bmatrix} C \\ Q \\ E \end{bmatrix} = \begin{bmatrix} 0 & 2 & -2 \\ -1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} I(X;B)_\sigma \\ \frac{1}{2} I(A;B|X)_\sigma \\ -\frac{1}{2} I(A;E|X)_\sigma \end{bmatrix}. \tag{24.41}$$

We can now determine bounds on an achievable rate region that employs the above coding strategy. We apply the inverse of the matrix in (24.34) to the LHS and RHS, giving

$$\begin{bmatrix} -\frac{1}{2} & -1 & 0 \\ 0 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} C \\ Q \\ E \end{bmatrix} - \begin{bmatrix} -\frac{1}{2} & -1 & 0 \\ 0 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} I(X;B)_\sigma \\ \frac{1}{2} I(A;B|X)_\sigma \\ -\frac{1}{2} I(A;E|X)_\sigma \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}. \tag{24.42}$$

Then using the following identities

$$I(X;B)_\sigma + I(A;B|X)_\sigma = I(AX;B)_\sigma, \tag{24.43}$$

$$\frac{1}{2} I(A;B|X)_\sigma - \frac{1}{2} I(A;E|X)_\sigma = I(A\rangle BX)_\sigma, \tag{24.44}$$

and the constraints $\alpha, \beta, \gamma \geq 0$, we obtain the inequalities in (24.8-24.10), corresponding exactly to the one-shot, one-state region in Theorem 24.2.1. Taking the union over all possible states σ in (24.11) and taking the regularization gives the full dynamic achievable rate region.

Figure 24.2 illustrates an example of the general polyhedron specified by (24.8-24.10), where the channel is the qubit dephasing channel $\rho \to (1-p)\rho + pZ\rho Z$ with dephasing parameter $p = 0.2$, and the input state is

$$\sigma^{XAA'} \equiv \frac{1}{2}(|0\rangle\langle 0|^X \otimes \phi_0^{AA'} + |1\rangle\langle 1|^X \otimes \phi_1^{AA'}), \tag{24.45}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where

$$|\phi_0\rangle^{AA'} \equiv \sqrt{1/4}|00\rangle^{AA'} + \sqrt{3/4}|11\rangle^{AA'}, \tag{24.46}$$

$$|\phi_1\rangle^{AA'} \equiv \sqrt{3/4}|00\rangle^{AA'} + \sqrt{1/4}|11\rangle^{AA'}. \tag{24.47}$$

The state σ^{XABE} resulting from the channel is $U_{\mathcal{N}}^{A'\to BE}(\sigma^{XAA'})$ where $U_{\mathcal{N}}$ is an isometric extension of the qubit dephasing channel. The figure caption provides a detailed explanation of the one-shot, one-state region $\mathcal{C}^{(1)}_{\text{CQE},\sigma}$ (note that Figure 24.2 displays the one-shot, one-state region and does not display the full capacity region).

24.4 The Converse Theorem

We provide a catalytic, information theoretic converse proof of the dynamic capacity region, showing that (24.6) gives a multi-letter characterization of it. The catalytic approach means that we are considering the most general protocol that *consumes and generates* classical communication, quantum communication, and entanglement in addition to the uses of the noisy quantum channel. This approach has the advantage that we can prove the converse theorem in "one fell swoop." We employ the Alicki-Fannes' inequality, the chain rule for quantum mutual information, elementary properties of quantum entropy, and the quantum data processing inequality to prove the converse.

We show that the bounds in (24.8-24.10) hold for common randomness generation instead of classical communication because a capacity for generating common randomness can only be better than that for generating classical communication (classical communication can generate common randomness). We also consider a protocol that preserves entanglement with a reference system instead of one that generates quantum communication.

We prove that the converse theorem holds for a state of the following form

$$\sigma^{XAB} \equiv \sum_x p(x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A'\to B}(\rho_x^{AA'}), \tag{24.48}$$

where the states $\rho_x^{AA'}$ are mixed, rather than proving it for a state of the form in (24.11). Then we show in Section 24.4.1 that it is not necessary to consider an ensemble of mixed states—i.e., we can do just as well with an ensemble of pure states, giving the statement of Theorem 24.2.1.

We first prove the bound in (24.8). Consider the following chain of inequalities:

$$\begin{aligned} n\big(\bar{C} + 2\bar{Q}\big) &= I(M;\hat{M})_{\overline{\Phi}} + I(R;B_1)_{\Phi} && (24.49)\\ &\leq I(M;\hat{M})_{\omega'} + I(R;B_1)_{\omega'} + n\delta' && (24.50)\\ &\leq I(M;B^nA_2LT_B)_{\omega} + I(R;B^nA_2LT_B)_{\omega} && (24.51)\\ &\leq I(M;B^nA_2LT_B)_{\omega} + I(R;B^nA_2LT_BM)_{\omega} && (24.52)\\ &= I(M;B^nA_2LT_B)_{\omega} + I(R;B^nA_2LT_B|M)_{\omega} + I(R;M)_{\omega}. && (24.53) \end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

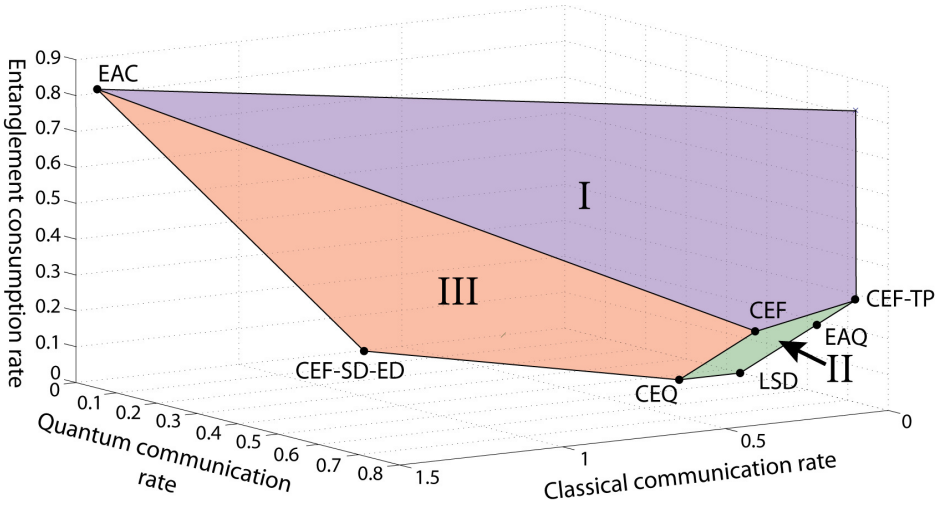

Figure 24.2: An example of the one-shot, one-state achievable region $\mathcal{C}^{(1)}_{\mathrm{CQE},\sigma}(\mathcal{N})$ corresponding to a state σ^{XABE} that arises from a qubit dephasing channel with dephasing parameter $p = 0.2$. The figure depicts the octant corresponding to the consumption of entanglement and the generation of classical and quantum communication. The state input to the channel $\mathcal{N}$ is $\sigma^{XAA'}$, defined in (24.45). The plot features seven achievable corner points of the one-shot, one-state region. We can achieve the convex hull of these seven points by time-sharing any two different coding strategies. We can also achieve any point above an achievable point by consuming more entanglement than necessary. The seven achievable points correspond to entanglement-assisted quantum communication (EAQ), the protocol from Corollary 21.5.3 for classically-enhanced quantum communication (CEQ), the protocol from Theorem 21.5.1 for entanglement-assisted classical communication with limited entanglement (EAC), quantum communication (LSD), combining CEF with entanglement distribution and super-dense coding (CEF-SD-ED), the protocol from Corollary 21.5.2 for entanglement-assisted communication of classical and quantum information (CEF), and combining CEF with teleportation (CEF-TP). Observe that we can obtain EAC by combining CEF with super-dense coding, so that the points CEQ, CEF, EAC, and CEF-SD-ED all lie in plane III. Observe that we can obtain CEQ from CEF by entanglement distribution and we can obtain LSD from EAQ and EAQ from CEF-TP, both by entanglement distribution. Thus, the points CEF, CEQ, LSD, EAQ, and CEF-TP all lie in plane II. Finally, observe that we can obtain all corner points by combining CEF with the unit protocols of teleportation, super-dense coding, and entanglement distribution. The bounds in (24.8-24.10) uniquely specify the respective planes I-III. We obtain the full achievable region by taking the union over all states σ of the one-shot, one-state regions $\mathcal{C}^{(1)}_{\sigma}(\mathcal{N})$ and taking the regularization, as outlined in Theorem 24.2.1. The above region is a translation of the unit resource capacity region from Chapter 8 to the protocol for entanglement-assisted communication of classical and quantum information.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first equality holds by evaluating the quantum mutual informations on the respective states $\overline{\Phi}^{M\hat{M}}$ and Φ^{RB_1}. The first inequality follows from the condition in (24.5) and an application of the Alicki-Fannes' inequality where δ' vanishes as $\epsilon \to 0$. We suppress this term in the rest of the inequalities for convenience. The second inequality follows from quantum data processing, and the third follows from another application of quantum data processing. The second equality follows by applying the mutual information chain rule. We continue below:

$$
\begin{aligned}
&= I(M; B^n A_2 L T_B)_\omega + I(R; B^n A_2 L T_B | M)_\omega && (24.54)\\
&= I(M; B^n A_2 L T_B)_\omega + I(R A_2 L T_B; B^n | M)_\omega \\
&\quad + I(R; A_2 L T_B | M)_\omega - I(B^n; A_2 L T_B | M)_\omega && (24.55)\\
&= I(M; B^n)_\omega + I(M; A_2 L T_B | B^n)_\omega + I(R A_2 L T_B; B^n | M)_\omega \\
&\quad + I(R; A_2 L T_B | M)_\omega - I(B^n; A_2 L T_B | M)_\omega && (24.56)
\end{aligned}
$$

$$
\begin{aligned}
&= I(R A_2 L M T_B; B^n)_\omega + I(M; A_2 L T_B | B^n)_\omega \\
&\quad + I(R; A_2 L T_B | M)_\omega - I(B^n; A_2 L T_B | M)_\omega && (24.57)\\
&\leq I(R A_2 T_B S_A L M; B^n)_\omega + I(M; A_2 L T_B | B^n)_\omega \\
&\quad + I(R; A_2 L T_B | M)_\omega - I(B^n; A_2 L T_B | M)_\omega && (24.58)\\
&= I(AX; B^n)_\omega + I(M; A_2 L T_B | B^n)_\omega \\
&\quad + I(R; A_2 L T_B | M)_\omega - I(B^n; A_2 L T_B | M)_\omega. && (24.59)
\end{aligned}
$$

The first equality follows because $I(R; M)_\omega = 0$ for this protocol. The second equality follows from applying the chain rule for quantum mutual information to the term $I(R; B^n A_2 L T_B | M)_\omega$, and the third is another application of the chain rule to the term $I(M; B^n A_2 L T_B)_\omega$. The fourth equality follows by combining $I(M; B^n)_\omega$ and $I(R A_2 L T_B; B^n | M)_\omega$ with the chain rule. The inequality follows from an application of quantum data processing. The final equality follows from the definitions $A \equiv R T_B A_2 S_A$ and $X \equiv ML$. We now focus on the term $I(M; A_2 L T_B | B^n)_\omega + I(R; A_2 L T_B | M)_\omega - I(B^n; A_2 L T_B | M)_\omega$ and show that it is less than $n\left(\tilde{C} + 2\tilde{Q}\right)$:

$$
\begin{aligned}
&I(M; A_2 L T_B | B^n)_\omega + I(R; A_2 L T_B | M)_\omega - I(B^n; A_2 L T_B | M)_\omega \\
&= I(M; A_2 L T_B B^n)_\omega + I(R; A_2 L T_B M)_\omega - I(B^n; A_2 L T_B M)_\omega - I(R; M)_\omega && (24.60)\\
&= I(M; A_2 L T_B B^n)_\omega + I(R; A_2 L T_B M)_\omega - I(B^n; A_2 L T_B M)_\omega && (24.61)\\
&= H(A_2 L T_B B^n)_\omega + H(R)_\omega - H(R A_2 L T_B | M)_\omega - H(B^n)_\omega && (24.62)\\
&= H(A_2 L T_B B^n)_\omega - H(A_2 L T_B | M R)_\omega - H(B^n)_\omega && (24.63)\\
&= H(A_2 L T_B | B^n)_\omega - H(A_2 L T_B | M R)_\omega. && (24.64)
\end{aligned}
$$

The first equality follows by applying the chain rule for quantum mutual information. The second equality follows because $I(R; M)_\omega = 0$ for this protocol. The third equality follows

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

by expanding the quantum mutual informations. The next two inequalities follow from straightforward entropic manipulations and that $H(R)_\omega = H(R|M)_\omega$ for this protocol. We continue below:

$$
\begin{aligned}
&= H(A_2L|B^n)_\omega + H(T_B|B^nA_2L)_\omega - H(T_B|MR)_\omega - H(A_2L|T_BMR)_\omega && (24.65)\\
&= H(A_2L|B^n)_\omega + H(T_B|B^nA_2L)_\omega - H(T_B)_\omega - H(A_2L|T_BMR)_\omega && (24.66)\\
&= H(A_2L|B^n)_\omega - I(T_B;B^nA_2L)_\omega - H(A_2L|T_BMR) && (24.67)\\
&\leq H(A_2L)_\omega - H(A_2L|T_BMR)_\omega && (24.68)\\
&= I(A_2L;T_BMR)_\omega && (24.69)\\
&= I(L;T_BMR)_\omega + I(A_2;T_BMR|L)_\omega && (24.70)\\
&\leq n\Big(\tilde{C} + 2\tilde{Q}\Big). && (24.71)
\end{aligned}
$$

The first two equalities follow from the chain rule for entropy and the second exploits that $H(T_B|MR) = H(T_B)$ for this protocol. The third equality follows from the definition of quantum mutual information. The inequality follows from subadditivity of entropy and that $I(T_B;B^nA_2L)_\omega \geq 0$. The fourth equality follows from the definition of quantum mutual information and the next equality follows from the chain rule. The final inequality follows because the quantum mutual information $I(L;T_BMR)_\omega$ can never be larger than the logarithm of the dimension of the classical register L and because the quantum mutual information $I(A_2;T_BMR|L)_\omega$ can never be larger than twice the logarithm of the dimension of the quantum register A_2. Thus the following inequality applies

$$
n\big(\bar{C} + 2\bar{Q}\big) \leq I(AX;B^n)_\omega + n\Big(\tilde{C} + 2\tilde{Q}\Big) + n\delta', \tag{24.72}
$$

demonstrating that (24.8) holds for the net rates.

We now prove the second bound in (24.9). Consider the following chain of inequalities:

$$
\begin{aligned}
n\big(\bar{Q} + \bar{E}\big) &= I(R\rangle B_1)_\Phi + I(S_A\rangle S_B)_\Phi && (24.73)\\
&= I(RS_A\rangle B_1S_B)_{\Phi\otimes\Phi} && (24.74)\\
&\leq I(RS_A\rangle B_1S_B)_{\omega'} + n\delta' && (24.75)\\
&\leq I(RS_A\rangle B_1S_BM)_{\omega'} && (24.76)\\
&\leq I(RS_A\rangle B^nA_2T_BLM)_\omega && (24.77)\\
&= H(B^nA_2T_B|LM)_\omega - H(RS_AB^nA_2T_B|LM)_\omega && (24.78)\\
&\leq H(B^n|LM)_\omega + H(A_2|LM)_\omega + H(T_B|LM)_\omega \\
&\qquad - H(RS_AB^nA_2T_B|LM)_\omega && (24.79)\\
&\leq I(RS_AA_2T_B\rangle B^nLM)_\omega + n\Big(\tilde{Q} + \tilde{E}\Big) && (24.80)\\
&= I(A\rangle B^nX)_\omega + n\Big(\tilde{Q} + \tilde{E}\Big). && (24.81)
\end{aligned}
$$

The first equality follows by evaluating the coherent informations of the respective states Φ^{RB_1} and $\Phi^{S_AS_B}$. The second equality follows because $\Phi^{RB_1}\otimes\Phi^{T_AT_B}$ is a product state. The

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

first inequality follows from the condition in (24.5) and an application of the Alicki-Fannes' inequality with δ' vanishing when $\epsilon \to 0$. We suppress the term $n\delta'$ in the following lines. The next two inequalities follow from quantum data processing. The third equality follows from the definition of coherent information. The fourth inequality follows from subadditivity of entropy. The fifth inequality follows from the definition of coherent information and the fact that the entropy can never be larger than the logarithm of the dimension of the corresponding system. The final equality follows from the definitions $A \equiv RT_BA_2S_A$ and $X \equiv ML$. Thus the following inequality applies

$$n\big(\bar{Q} + \bar{E}\big) \leq I(A\rangle B^n X) + n\Big(\tilde{Q} + \tilde{E}\Big), \tag{24.82}$$

demonstrating that (24.9) holds for the net rates.

We prove the last bound in (24.10). Consider the following chain of inequalities:

$$\begin{aligned}
n\big(\bar{C} + \bar{Q} + \bar{E}\big) &= I(M;\hat{M})_{\overline{\Phi}} + I(RS_A\rangle B_1S_B)_{\Phi\otimes\Phi} && (24.83)\\
&\leq I(M;\hat{M})_{\omega'} + I(RS_A\rangle B_1S_B)_{\omega'} + n\delta' && (24.84)\\
&\leq I(M;B^nA_2T_BL)_\omega + I(RS_A\rangle B^nA_2T_BLM)_\omega && (24.85)\\
&= I(ML;B^nA_2T_B)_\omega + I(M;L)_\omega - I(A_2B^nT_B;L)_\omega \\
&\quad + H(B^n|LM) + H(A_2T_B|B^nLM)_\omega \\
&\quad - H(RS_AA_2T_BB^n|LM)_\omega && (24.86)\\
&= I(ML;B^n)_\omega + I(ML;A_2T_B|B^n)_\omega \\
&\quad + I(M;L)_\omega - I(A_2B^nT_B;L)_\omega \\
&\quad + H(A_2T_B|B^nLM)_\omega + I(RS_AA_2T_B\rangle B^nLM)_\omega. && (24.87)
\end{aligned}$$

The first equality follows from evaluating the mutual information of the state $\overline{\Phi}^{M\hat{M}}$ and the coherent information of the product state $\Phi^{RB_1} \otimes \Phi^{S_AS_B}$. The first inequality follows from the condition in (24.5) and an application of the Alicki-Fannes' inequality with δ' vanishing when $\epsilon \to 0$. We suppress the term $n\delta'$ in the following lines. The second inequality follows from quantum data processing. The second equality follows from applying the chain rule for quantum mutual information to $I(M;B^nA_2T_BL)_\omega$ and by expanding the coherent information $I(RS_A\rangle B^nA_2T_BLM)_\omega$. The third equality follows from applying the chain rule for quantum mutual information to $I(ML;B^nA_2T_B)_\omega$ and from the definition of coherent information. We continue below:

$$\begin{aligned}
&= I(ML;B^n)_\omega + I(RS_AA_2T_B\rangle B^nLM)_\omega \\
&\quad + I(ML;A_2T_B|B^n)_\omega + I(M;L)_\omega \\
&\quad - I(A_2B^nT_B;L)_\omega + H(A_2T_B|B^nLM)_\omega && (24.88)\\
&= I(ML;B^n)_\omega + I(RS_AA_2T_B\rangle B^nLM)_\omega \\
&\quad + H(A_2T_B|B^n)_\omega + I(M;L)_\omega - I(A_2B^nT_B;L)_\omega && (24.89)
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$$\leq I(ML;B^n)_\omega + I(RS_AA_2T_B\rangle B^nLM)_\omega + n\Big(\tilde{C}+\tilde{Q}+\tilde{E}\Big) \tag{24.90}$$

$$= I(X;B^n)_\omega + I(A\rangle B^nX)_\omega + n\Big(\tilde{C}+\tilde{Q}+\tilde{E}\Big). \tag{24.91}$$

The first equality follows by rearranging terms. The second equality follows by canceling terms. The inequality follows from the fact that the entropy $H(A_2T_B|B^n)_\omega$ can never be larger than the logarithm of the dimension of the systems A_2T_B, that the mutual information $I(M;L)_\omega$ can never be larger than the logarithm of the dimension of the classical register L, and because $I(A_2B^nT_B;L)_\omega \geq 0$. The last equality follows from the definitions $A \equiv RT_BA_2S_A$ and $X \equiv ML$. Thus the following inequality holds

$$n\big(\bar{C}+\bar{Q}+\bar{E}\big) \leq I(X;B^n)_\omega + I(A\rangle B^nX)_\omega + n\Big(\tilde{C}+\tilde{Q}+\tilde{E}\Big) + n\delta', \tag{24.92}$$

demonstrating that the inequality in (24.10) applies to the net rates. This concludes the catalytic proof of the converse theorem.

24.4.1 Pure state ensembles are sufficient

We prove that it is sufficient to consider an ensemble of pure states as in the statement of Theorem 24.2.1 rather than an ensemble of mixed states as in (24.48) in the proof of our converse theorem. We first determine a spectral decomposition of the mixed state ensemble, model the index of the pure states in the decomposition as a classical variable Y, and then place this classical variable Y in a classical register. It follows that the communication rates can only improve, and it is sufficient to consider an ensemble of pure states.

Consider that each mixed state in the ensemble in (24.48) admits a spectral decomposition of the following form:

$$\rho_x^{AA'} = \sum_y p(y|x)\psi_{x,y}^{AA'}. \tag{24.93}$$

We can thus represent the ensemble as follows:

$$\rho^{XAB} \equiv \sum_{x,y} p(x)p(y|x)|x\rangle\langle x|^X \otimes \mathcal{N}^{A'\to B}(\psi_{x,y}^{AA'}). \tag{24.94}$$

The inequalities in (24.8-24.10) for the dynamic capacity region involve the mutual information $I(AX;B)_\rho$, the Holevo information $I(X;B)_\rho$, and the coherent information $I(A\rangle BX)_\rho$. As we show below, each of these entropic quantities can only improve in each case if we make the variable y be part of the classical variable. This improvement then implies that it is only necessary to consider pure states in the dynamic capacity theorem.

Let θ^{XYAB} denote an augmented state of the following form:

$$\theta^{XYAB} \equiv \sum_x p(x)p(y|x)|x\rangle\langle x|^X \otimes |y\rangle\langle y|^Y \otimes \mathcal{N}^{A'\to B}(\psi_{x,y}^{AA'}). \tag{24.95}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

This state is actually a state of the form in (24.11) if we subsume the classical variables X and Y into one classical variable. The following three inequalities each follow from an application of the quantum data processing inequality:

$$I(X;B)_\rho = I(X;B)_\theta \leq I(XY;B)_\theta, \tag{24.96}$$

$$I(AX;B)_\rho = I(AX;B)_\theta \leq I(AXY;B)_\theta \tag{24.97}$$

$$I(A\rangle BX)_\rho = I(A\rangle BX)_\theta \leq I(A\rangle BXY)_\theta. \tag{24.98}$$

Each of these inequalities proves the desired result for the respective Holevo information, mutual information, and coherent information, and it suffices to consider an ensemble of pure states in Theorem 24.2.1.

24.4.2 The Quantum Dynamic Capacity Formula

We introduce the quantum dynamic capacity formula and show how additivity of it implies that the computation of the Pareto optimal trade-off surface of the capacity region requires an optimization over a single channel use, rather than an infinite number of them. The Pareto optimal trade-off surface consists of all points in the capacity region that are Pareto optimal, in the sense that it is not possible to make improvements in one resource without offsetting another resource (these are essentially the boundary points of the region in our case). We then show how several important capacity formulas discussed previously in this book are special cases of the quantum dynamic capacity formula.

Definition 24.4.1 (Quantum Dynamic Capacity Formula)**.** *The quantum dynamic capacity formula of a quantum channel $\mathcal{N}$ is as follows:*

$$D_{\lambda,\mu}(\mathcal{N}) \equiv \max_\sigma I(AX;B)_\sigma + \lambda I(A\rangle BX)_\sigma + \mu(I(X;B)_\sigma + I(A\rangle BX)_\sigma), \tag{24.99}$$

where σ is a state of the form in (24.11), $\lambda, \mu \geq 0$, and these parameters λ and μ play the role of Lagrange multipliers.

Definition 24.4.2. *The regularized quantum dynamic capacity formula is as follows:*

$$D^{reg}_{\lambda,\mu}(\mathcal{N}) \equiv \lim_{k\to\infty} \frac{1}{k} D_{\lambda,\mu}\big(\mathcal{N}^{\otimes k}\big). \tag{24.100}$$

Lemma 24.4.1. *Suppose the quantum dynamic capacity formula is additive for a channel $\mathcal{N}$ and any other arbitrary channel $\mathcal{M}$:*

$$D_{\lambda,\mu}(\mathcal{N}\otimes\mathcal{M}) = D_{\lambda,\mu}(\mathcal{N}) + D_{\lambda,\mu}(\mathcal{M}). \tag{24.101}$$

Then the regularized quantum dynamic capacity formula for $\mathcal{N}$ is equal to the quantum dynamic capacity formula:

$$D^{reg}_{\lambda,\mu}(\mathcal{N}) = D_{\lambda,\mu}(\mathcal{N}). \tag{24.102}$$

In this sense, the regularized formula "single-letterizes" and it is not necessary to take the limit.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We prove the result using induction on n. The base case for $n = 1$ is trivial. Suppose the result holds for n: $D_{\lambda,\mu}(\mathcal{N}^{\otimes n}) = nD_{\lambda,\mu}(\mathcal{N})$. Then the following chain of equalities proves the inductive step:

$$\begin{aligned} D_{\lambda,\mu}(\mathcal{N}^{\otimes n+1}) &= D_{\lambda,\mu}(\mathcal{N} \otimes \mathcal{N}^{\otimes n}) && (24.103)\\ &= D_{\lambda,\mu}(\mathcal{N}) + D_{\lambda,\mu}(\mathcal{N}^{\otimes n}) && (24.104)\\ &= D_{\lambda,\mu}(\mathcal{N}) + nD_{\lambda,\mu}(\mathcal{N}). && (24.105) \end{aligned}$$

The first equality follows by expanding the tensor product. The second critical equality follows from the assumption that the formula is additive. The final equality follows from the induction hypothesis.

Theorem 24.4.1. *Single-letterization of the quantum dynamic capacity formula implies that the computation of the Pareto optimal trade-off surface of the dynamic capacity region requires an optimization over a single channel use.*

We employ ideas from optimization theory for the proof (see Ref. [45]). We would like to characterize all the points in the capacity region that are Pareto optimal. Such a task is standard vector optimization in the theory of Pareto trade-off analysis (see Section 4.7 of Ref. [45]). We can phrase the optimization task as the following scalarization of the vector optimization task:

$$\max_{C,Q,E,p(x),\phi_x} w_C C + w_Q Q + w_E E \tag{24.106}$$

subject to

$$\begin{aligned} C + 2Q &\leq I(AX;B^n)_\sigma, && (24.107)\\ Q + E &\leq I(A\rangle B^n X)_\sigma, && (24.108)\\ C + Q + E &\leq I(X;B^n)_\sigma + I(A\rangle B^n X)_\sigma, && (24.109) \end{aligned}$$

where the maximization is over all C, Q, and E and over probability distributions $p_X(x)$ and bipartite states $\phi_x^{AA'^n}$. The geometric interpretation of the scalarization task is that we are trying to find a supporting plane of the dynamic capacity region where the weight vector (w_C, w_Q, w_E) is the normal vector of the plane and the value of its inner product with (C, Q, E) characterizes the offset of the plane.

The Lagrangian of the above optimization problem is

$$\begin{aligned} \mathcal{L}\Big(C,Q,E,p_X(x),\phi_x^{AA'^n},\lambda_1,\lambda_2,\lambda_3\Big) \equiv & \\ w_C C + w_Q Q + w_E E + \lambda_1(I(AX;B^n)_\sigma - (C+2Q)) & \\ + \lambda_2(I(A\rangle B^n X)_\sigma - (Q+E)) & \\ + \lambda_3(I(X;B^n)_\sigma + I(A\rangle B^n X)_\sigma - (C+Q+E)), & \quad (24.110) \end{aligned}$$

and the Lagrange dual function g [45] is

$$g(\lambda_1,\lambda_2,\lambda_3) \equiv \sup_{C,Q,E,p(x),\phi_x^{AA'^n}} \mathcal{L}\Big(C,Q,E,p_X(x),\phi_x^{AA'^n},\lambda_1,\lambda_2,\lambda_3\Big), \tag{24.111}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $\lambda_1, \lambda_2, \lambda_3 \geq 0$. The optimization task simplifies if the Lagrange dual function does. Thus, we rewrite the Lagrange dual function as follows:

$$\begin{aligned} & g(\lambda_1, \lambda_2, \lambda_3) \\ & = \sup_{C,Q,E,p(x),\phi_x^{AA'^n}} w_C C + w_Q Q + w_E E + \lambda_1 (I(AX;B^n)_\sigma - (C + 2Q)) \qquad (24.112) \\ & \quad + \lambda_2 (I(A\rangle B^n X)_\sigma - (Q + E)) \\ & \quad + \lambda_3 (I(X;B^n)_\sigma + I(A\rangle B^n X)_\sigma - (C + Q + E)) \qquad (24.113) \end{aligned}$$

$$\begin{aligned} & = \sup_{C,Q,E,p(x),\phi_x^{AA'^n}} (w_C - \lambda_1 - \lambda_3) C + (w_Q - 2\lambda_1 - \lambda_2 - \lambda_3) Q + (w_E - \lambda_2 - \lambda_3) E \\ & \quad + \lambda_1 \left(I(AX;B^n)_\sigma + \frac{\lambda_2}{\lambda_1} I(A\rangle B^n X)_\sigma + \frac{\lambda_3}{\lambda_1} (I(X;B^n)_\sigma + I(A\rangle B^n X)_\sigma) \right) \qquad (24.114) \\ & = \sup_{C,Q,E} (w_C - \lambda_1 - \lambda_3) C + (w_Q - 2\lambda_1 - \lambda_2 - \lambda_3) Q + (w_E - \lambda_2 - \lambda_3) E \\ & \quad + \lambda_1 \left(\max_{p(x),\phi_x^{AA'^n}} I(AX;B^n)_\sigma + \frac{\lambda_2}{\lambda_1} I(A\rangle B^n X)_\sigma + \frac{\lambda_3}{\lambda_1} (I(X;B^n)_\sigma + I(A\rangle B^n X)_\sigma) \right). \\ & \qquad (24.115) \end{aligned}$$

The first equality follows by definition. The second equality follows from some algebra, and the last follows because the Lagrange dual function factors into two separate optimization tasks: one over C, Q, and E and another that is equivalent to the quantum dynamic capacity formula with $\lambda = \lambda_2/\lambda_1$ and $\mu = \lambda_3/\lambda_1$. Thus, the computation of the Pareto optimal trade-off surface requires just a single use of the channel if the quantum dynamic capacity formula in (24.99) single-letterizes.

Special cases of the quantum dynamic capacity formula

We now show how several capacity formulas of a quantum channel, including the entanglement-assisted classical capacity (Theorem 20.3.1), the quantum capacity formula (Theorem 23.3.1), and the classical capacity formula (Theorem 19.3.1) are special cases of the quantum dynamic capacity formula.

We first give a geometric interpretation of these special cases before proceeding to the proofs. Recall that the dynamic capacity region has the simple interpretation as a translation of the three-faced unit resource capacity region along the trade-off curve for entanglement-assisted classical and quantum communication (see Figure 24.4 for the example of the region of the dephasing channel). Any particular weight vector (w_C, w_Q, w_E) in (24.106) gives a set of parallel planes that slice through the (C, Q, E) space, and the goal of the scalar optimization task is to find one of these planes that is a supporting plane, intersecting a point (or a set of points) on the trade-off surface of the dynamic capacity region. We consider three special planes:

1. The first corresponds to the plane containing the vectors of super-dense coding and teleportation. The normal vector of this plane is $(1, 2, 0)$, and suppose that we set the

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

weight vector in (24.106) to be this vector. Then the optimization program finds the set of points on the trade-off surface such that a plane with this normal vector is a supporting plane for the region. The optimization program singles out (24.107), and we can think of this as being equivalent to setting $\lambda_2, \lambda_3 = 0$ in the Lagrange dual function. We show below that the optimization program becomes equivalent to finding the entanglement-assisted capacity (Theorem 20.3.1), in the sense that the quantum dynamic capacity formula becomes the entanglement-assisted capacity formula.

2. The next plane contains the vectors of teleportation and entanglement distribution. The normal vector of this plane is $(0, 1, 1)$. Setting the weight vector in (24.106) to be this vector makes the optimization program single out (24.108), and we can think of this as being equivalent to setting $\lambda_1, \lambda_3 = 0$ in the Lagrange dual function. We show below that the optimization program becomes equivalent to finding the quantum capacity (Theorem 23.3.1), in the sense that the quantum dynamic capacity formula becomes the LSD formula for the quantum capacity.

3. A third plane contains the vectors of super-dense coding and entanglement distribution. The normal vector of this plane is $(1, 1, 1)$. Setting the weight vector in (24.106) to be this vector makes the optimization program single out (24.109), and we can think of this as being equivalent to setting $\lambda_1, \lambda_2 = 0$ in the Lagrange dual function. We show below that the optimization becomes equivalent to finding the classical capacity (Theorem 19.3.1), in the sense that the quantum dynamic capacity formula becomes the HSW formula for the classical capacity.

Corollary 24.4.1. *The quantum dynamic capacity formula is equivalent to the entanglement-assisted classical capacity formula when $\lambda, \mu = 0$, in the sense that*

$$\max_{\sigma} I(AX; B) = \max_{\phi^{AA'}} I(A; B). \tag{24.116}$$

Proof. The inequality $\max_\sigma I(AX; B) \geq \max_{\phi^{AA'}} I(A; B)$ follows because the state σ is of the form in (24.11) and we can always choose $p_X(x) = \delta_{x,x_0}$ and $\phi_{x_0}^{AA'}$ to be the state that maximizes $I(A; B)$. We now show the other inequality $\max_\sigma I(AX; B) \leq \max_{\phi^{AA'}} I(A; B)$. First, consider that the following chain of equalities holds for any state ϕ^{ABE} resulting from the isometric extension of the channel:

$$\begin{aligned} I(A; B) &= H(B) + H(A) - H(AB) && (24.117)\\ &= H(B) + H(BE) - H(E) && (24.118)\\ &= H(B) + H(B|E). \end{aligned}$$

In this way, we see that the mutual information is purely a function of the channel input density operator $\text{Tr}_A\{\phi^{AA'}\}$. Then consider any state σ of the form in (24.11). The following

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

chain of inequalities holds

$$\begin{aligned} I(AX;B)_\sigma &= H(A|X)_\sigma + H(B)_\sigma - H(E|X)_\sigma && (24.119)\\ &= H(BE|X)_\sigma + H(B)_\sigma - H(E|X)_\sigma && (24.120)\\ &= H(B|EX)_\sigma + H(B)_\sigma && (24.121)\\ &\leq H(B|E)_\sigma + H(B)_\sigma && (24.122)\\ &\leq \max_{\phi^{AA'}} I(A;B). && (24.123) \end{aligned}$$

The first equality follows by expanding the mutual information. The second equality follows because the state on ABE is pure when conditioned on X. The third equality follows from the entropy chain rule. The first inequality follows from strong subadditivity, and the last follows because the state after tracing out systems X and A is a particular state that arises from the channel and cannot be larger than the maximum. □

Corollary 24.4.2. *The quantum dynamic capacity formula is equivalent to the LSD quantum capacity formula in the limit where $\lambda \to \infty$ and μ is fixed, in the sense that*

$$\max_\sigma I(A\rangle BX) = \max_{\phi^{AA'}} I(A\rangle B). \tag{24.124}$$

Proof. The inequality $\max_\sigma I(A\rangle BX) \geq \max_{\phi^{AA'}} I(A\rangle B)$ follows because the state σ is of the form in (24.11) and we can always choose $p_X(x) = \delta_{x,x_0}$ and $\phi^{AA'}_{x_0}$ to be the state that maximizes $I(A\rangle B)$. The inequality $\max_\sigma I(A\rangle BX) \leq \max_{\phi^{AA'}} I(A\rangle B)$ follows because $I(A\rangle BX) = \sum_x p_X(x) I(A\rangle B)_{\phi_x}$ and the maximum is always greater than the average. □

Corollary 24.4.3. *The quantum dynamic capacity formula is equivalent to the HSW classical capacity formula in the limit where $\mu \to \infty$ and λ is fixed, in the sense that*

$$\max_\sigma I(A\rangle BX)_\sigma + I(X;B)_\sigma = \max_{\{p_X(x),\psi_x\}} I(X;B). \tag{24.125}$$

Proof. The inequality $\max_\sigma I(A\rangle BX)_\sigma + I(X;B)_\sigma \geq \max_{\{p_X(x),\psi_x\}} I(X;B)$ follows by choosing σ to be the pure ensemble that maximizes $I(X;B)$ and noting that $I(A\rangle BX)_\sigma$ vanishes for a pure ensemble. We now prove the inequality $\max_\sigma I(A\rangle BX)_\sigma + I(X;B)_\sigma \leq \max_{\{p_X(x),\psi_x\}} I(X;B)$. Consider a state ω^{XYBE} obtained by performing a von Neumann measurement on the A system of the state σ^{XABE}. Then

$$\begin{aligned} I(A\rangle BX)_\sigma + I(X;B)_\sigma &= H(B)_\sigma - H(E|X)_\sigma && (24.126)\\ &= H(B)_\omega - H(E|X)_\omega && (24.127)\\ &\leq H(B)_\omega - H(E|XY)_\omega && (24.128) \end{aligned}$$

$$\begin{aligned} &= H(B)_\omega - H(B|XY)_\omega && (24.129)\\ &= I(XY;B)_\omega && (24.130)\\ &\leq \max_{\{p_X(x),\psi_x\}} I(X;B). && (24.131) \end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The first equality follows by expanding the conditional coherent information and the Holevo information. The second equality follows because the measured A system is not involved in the entropies. The first inequality follows because conditioning does not increase entropy. The third equality follows because the state ω is pure when conditioned on X and Y. The fourth equality follows by definition, and the last inequality follows for clear reasons. □

24.5 Examples of Channels

In this final section, we prove that a broad class of channels, known as the Hadamard channels (see Section 5.2.4), have a single-letter dynamic capacity region. We prove this result by analyzing the quantum dynamic capacity formula for this class of channels. A dephasing channel is a special case of a Hadamard channel, and so we can compute its dynamic capacity region.

We also overview the dynamic capacity region of a lossy bosonic channel, which is a good model for free-space communication or loss in an optical fiber. Though, we only state the main results and do not get into too many details of this channel (doing so requires the theory of quantum optics and infinite-dimensional Hilbert spaces which is beyond the scope of this book). The upshot for this channel is that trade-off coding can give remarkable gains over time-sharing.

24.5.1 Quantum Hadamard channels

Below we show that the regularization in (24.6) is not necessary if the quantum channel is a Hadamard channel. This result holds because a Hadamard channel has a special structure (see Section 5.2.4).

Theorem 24.5.1. *The dynamic capacity region $\mathcal{C}_{\text{CQE}}(\mathcal{N}_H)$ of a quantum Hadamard channel $\mathcal{N}_H$ is equal to its one-shot region $\mathcal{C}^{(1)}_{\text{CQE}}(\mathcal{N}_H)$.*

The proof of the above theorem follows in two parts: 1) the below lemma shows the quantum dynamic capacity formula is additive when one of the channels is Hadamard and 2) the induction argument in Lemma 24.4.1 that proves single-letterization.

Lemma 24.5.1. *The following additivity relation holds for a Hadamard channel $\mathcal{N}_H$ and any other channel $\mathcal{N}$:*

$$D_{\lambda,\mu}(\mathcal{N}_H \otimes \mathcal{N}) = D_{\lambda,\mu}(\mathcal{N}_H) + D_{\lambda,\mu}(\mathcal{N}). \tag{24.132}$$

We first note that the inequality $D_{\lambda,\mu}(\mathcal{N}_H \otimes \mathcal{N}) \geq D_{\lambda,\mu}(\mathcal{N}_H) + D_{\lambda,\mu}(\mathcal{N})$ holds for any two channels simply by selecting the state σ in the maximization to be a tensor product of the ones that individually maximize $D_{\lambda,\mu}(\mathcal{N}_H)$ and $D_{\lambda,\mu}(\mathcal{N})$.

So we prove that the non-trivial inequality $D_{\lambda,\mu}(\mathcal{N}_H \otimes \mathcal{N}) \leq D_{\lambda,\mu}(\mathcal{N}_H) + D_{\lambda,\mu}(\mathcal{N})$ holds when the first channel is a Hadamard channel. Since the first channel is Hadamard, it is degradable and its degrading map has a particular structure: there are maps $\mathcal{D}_1^{B_1\to Y}$ and

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

$\mathcal{D}_2^{Y\to E_1}$ where Y is a classical register and such that the degrading map is $\mathcal{D}_2^{Y\to E_1}\circ\mathcal{D}_1^{B_1\to Y}$. Suppose the state we are considering to input to the tensor product channel is

$$\rho^{XAA_1'A_2'} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes \phi_x^{AA_1'A_2'}, \tag{24.133}$$

and this state is the one that maximizes $D_{\lambda,\mu}(\mathcal{N}_H\otimes\mathcal{N})$. Suppose that the output of the first channel is

$$\theta^{XAB_1E_1A_2'} \equiv U_{\mathcal{N}_H}^{A_1'\to B_1E_1}(\rho^{XAA_1'A_2'}), \tag{24.134}$$

and the output of the second channel is

$$\omega^{XAB_1E_1B_2E_2} \equiv U_{\mathcal{N}}^{A_2'\to B_2E_2}(\theta^{XAB_1E_1A_2'}). \tag{24.135}$$

Finally, we define the following state as the result of applying the first part of the Hadamard degrading map (a von Neumann measurement) to ω:

$$\sigma^{XYAE_1B_2E_2} \equiv \mathcal{D}_1^{B_1\to Y}(\omega^{XAB_1E_1B_2E_2}). \tag{24.136}$$

In particular, the state σ on systems $AE_1B_2E_2$ is pure when conditioned on X and Y. Then the following chain of inequalities holds

$$\begin{aligned}
&D_{\lambda,\mu}(\mathcal{N}_H\otimes\mathcal{N})\\
&= I(AX;B_1B_2)_\omega + \lambda I(A\rangle B_1B_2X)_\omega + \mu(I(X;B_1B_2)_\omega + I(A\rangle B_1B_2X)_\omega) && (24.137)\\
&= H(B_1B_2E_1E_2|X)_\omega + \lambda H(B_1B_2|X)_\omega + (\mu+1)H(B_1B_2)_\omega\\
&\quad - (\lambda+\mu+1)H(E_1E_2|X)_\omega && (24.138)\\
&= H(B_1E_1|X)_\omega + \lambda H(B_1|X)_\omega + (\mu+1)H(B_1)_\omega - (\lambda+\mu+1)H(E_1|X)_\omega +\\
&\quad H(B_2E_2|B_1E_1X)_\omega + \lambda H(B_2|B_1X)_\omega + (\mu+1)H(B_2|B_1)_\omega\\
&\quad - (\lambda+\mu+1)H(E_2|E_1X)_\omega && (24.139)
\end{aligned}$$

$$\begin{aligned}
&\leq H(B_1E_1|X)_\theta + \lambda H(B_1|X)_\theta + (\mu+1)H(B_1)_\theta - (\lambda+\mu+1)H(E_1|X)_\theta +\\
&\quad H(B_2E_2|YX)_\sigma + \lambda H(B_2|YX)_\sigma + (\mu+1)H(B_2)_\sigma - (\lambda+\mu+1)H(E_2|YX)_\sigma && (24.140)\\
&= I(AA_2'X;B_1)_\theta + \lambda I(AA_2'\rangle B_1X)_\theta + \mu(I(X;B_1)_\theta + I(AA_2'\rangle B_1X)_\theta) +\\
&\quad I(AE_1YX;B_2)_\sigma + \lambda I(AE_1\rangle B_2YX)_\sigma + \mu(I(YX;B_2)_\sigma + I(AE_1\rangle B_2YX)_\sigma) && (24.141)\\
&\leq D_{\lambda,\mu}(\mathcal{N}_H) + D_{\lambda,\mu}(\mathcal{N}). && (24.142)
\end{aligned}$$

The first equality follows by evaluating the quantum dynamic capacity formula $D_{\lambda,\mu}(\mathcal{N}_H\otimes\mathcal{N})$ on the state ρ. The next two equalities follow by rearranging entropies and because the state ω on systems $AB_1E_1B_2E_2$ is pure when conditioned on X. The inequality in the middle is the crucial one and follows from the Hadamard structure of the channel: we exploit monotonicity of conditional entropy under quantum operations so that $H(B_2|B_1X)_\omega \leq H(B_2|YX)_\sigma$, $H(B_2E_2|B_1E_1X)_\omega \leq H(B_2E_2|YX)_\sigma$, and $H(E_2|YX)_\sigma \leq H(E_2|E_1X)_\omega$. It also follows because $H(B_2|B_1)_\omega \leq H(B_2)_\omega$. The next equality follows by rearranging entropies and the final inequality follows because θ is a state of the form (24.11) for the first channel while σ is a state of the form (24.11) for the second channel.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

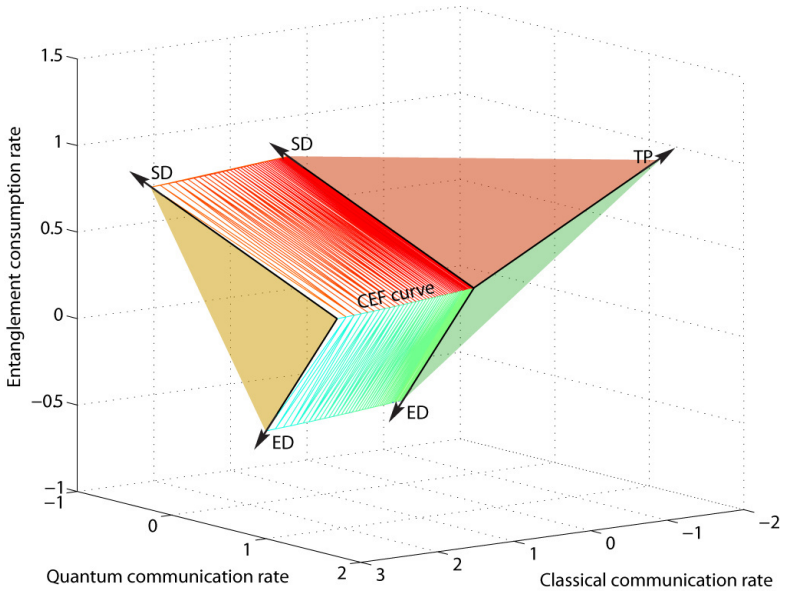

Figure 24.3: A plot of the dynamic capacity region for a qubit dephasing channel with dephasing parameter $p = 0.2$. The plot shows that the CEF trade-off curve (the protocol from Corollary 21.5.2) lies along the boundary of the dynamic capacity region. The rest of the region is simply the combination of the CEF points with the unit protocols teleportation (TP), super-dense coding (SD), and entanglement distribution (ED).

24.5.2 The Dephasing Channel

The below theorem shows that the full dynamic capacity region admits a particularly simple form when the noisy quantum channel is a qubit dephasing channel $\overline{\Delta}_p$ where

$$\overline{\Delta}_p(\rho) \equiv (1-p)\rho + p\overline{\Delta}(\rho), \tag{24.143}$$

$$\overline{\Delta}(\rho) \equiv \langle 0|\rho|0\rangle|0\rangle\langle 0| + \langle 1|\rho|1\rangle|1\rangle\langle 1|. \tag{24.144}$$

Figure 24.3 plots this region for the case of a dephasing channel with dephasing parameter $p = 0.2$. Figure 24.4 plots special two-dimensional cases of the full region for various values of the dephasing parameter p. The figure demonstrates that trade-off coding just barely beats time-sharing.

Theorem 24.5.2. *The dynamic capacity region $\mathcal{C}_{CQE}(\overline{\Delta}_p)$ of a dephasing channel with dephasing parameter p is the set of all C, Q, and E such that*

$$C + 2Q \leq 1 + H_2(\nu) - H_2(\gamma(\nu, p)), \tag{24.145}$$

$$Q + E \leq H_2(\nu) - H_2(\gamma(\nu, p)), \tag{24.146}$$

$$C + Q + E \leq 1 - H_2(\gamma(\nu, p)), \tag{24.147}$$

where $\nu \in [0, 1/2]$, H_2 is the binary entropy function, and

$$\gamma(\nu, p) \equiv \frac{1}{2} + \frac{1}{2}\sqrt{1 - 16 \cdot \frac{p}{2}\left(1 - \frac{p}{2}\right)\nu(1-\nu)}. \tag{24.148}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

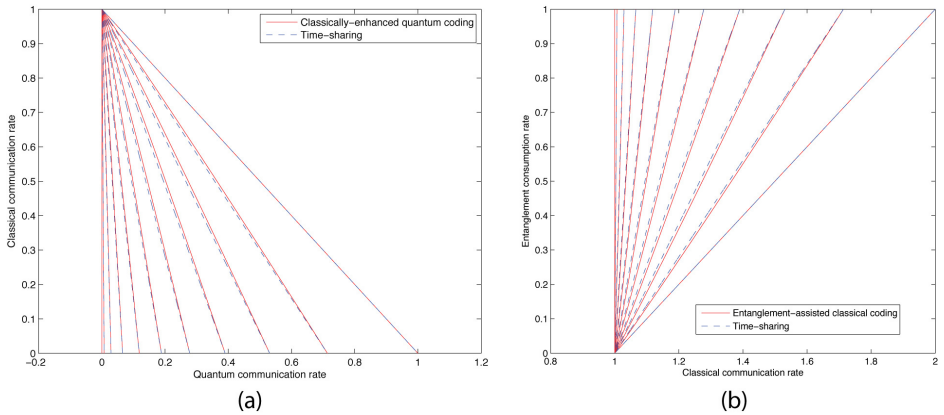

Figure 24.4: Plot of (a) the CQ trade-off curve and (b) the CE trade-off curve for a p-dephasing qubit channel for $p = 0, 0.1, 0.2, \ldots, 0.9, 1$. The trade-off curves for $p = 0$ correspond to those of a noiseless qubit channel and are the rightmost trade-off curve in each plot. The trade-off curves for $p = 1$ correspond to those for a classical channel, and are the leftmost trade-off curves in each plot. Each trade-off curve between these two extremes beats a time-sharing strategy, but these two extremes do not beat time-sharing.

We first notice that it suffices to consider an ensemble of pure states whose reductions to A' are diagonal in the dephasing basis (see the following exercise).

Exercise 24.5.1 Prove that the following properties hold for a generalized dephasing channel $\mathcal{N}_D$, its complement $\mathcal{N}_D^c$, the completely dephasing channel $\overline{\Delta}$, and all input states ρ:

$$\mathcal{N}_D\big(\overline{\Delta}(\rho)\big) = \overline{\Delta}(\mathcal{N}_D(\rho)), \tag{24.149}$$

$$\mathcal{N}_D^c\big(\overline{\Delta}(\rho)\big) = \mathcal{N}_D^c(\rho). \tag{24.150}$$

Conclude that

$$H(\rho) \leq H\big(\overline{\Delta}(\rho)\big), \tag{24.151}$$

$$H(\mathcal{N}_D(\rho)) \leq H\big(\overline{\Delta}(\mathcal{N}_D(\rho))\big) = H\big(\mathcal{N}_D\big(\overline{\Delta}(\rho)\big)\big), \tag{24.152}$$

$$H(\mathcal{N}_D^c(\rho)) = H\big(\mathcal{N}_D^c\big(\overline{\Delta}(\rho)\big)\big), \tag{24.153}$$

so that it suffices to consider diagonal input states for the dephasing channel.

Next we prove below that it is sufficient to consider an ensemble of the following form to characterize the boundary points of the region:

$$\frac{1}{2}|0\rangle\langle 0|^X \otimes \psi_0^{AA'} + \frac{1}{2}|1\rangle\langle 1|^X \otimes \psi_1^{AA'}, \tag{24.154}$$

where $\psi_0^{AA'}$ and $\psi_1^{AA'}$ are pure states, defined as follows for $\nu \in [0, 1/2]$:

$$\text{Tr}_A\left\{\psi_0^{AA'}\right\} = \nu|0\rangle\langle 0|^{A'} + (1-\nu)|1\rangle\langle 1|^{A'}, \tag{24.155}$$

$$\text{Tr}_A\left\{\psi_1^{AA'}\right\} = (1-\nu)|0\rangle\langle 0|^{A'} + \nu|1\rangle\langle 1|^{A'}. \tag{24.156}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We now prove the above claim. We assume without loss of generality that the dephasing basis is the computational basis. Consider a classical-quantum state with a finite number N of conditional density operators $\phi_x^{AA'}$ whose reduction to A' is diagonal:

$$\rho^{XAA'} \equiv \sum_{x=0}^{N-1} p_X(x)|x\rangle\langle x|^X \otimes \phi_x^{AA'}. \tag{24.157}$$

We can form a new classical-quantum state with double the number of conditional density operators by "bit-flipping" the original conditional density operators:

$$\sigma^{XAA'} \equiv \frac{1}{2}\sum_{x=0}^{N-1} p_X(x)\Big(|x\rangle\langle x|^X \otimes \phi_x^{AA'} + |x+N\rangle\langle x+N|^X \otimes X^{A'}\phi_x^{AA'}X^{A'}\Big), \tag{24.158}$$

where X is the σ_X "bit-flip" Pauli operator. Consider the following chain of inequalities that holds for all $\lambda, \mu \geq 0$:

$$\begin{aligned}
&I(AX;B)_\rho + \lambda I(A\rangle BX)_\rho + \mu\Big(I(X;B)_\rho + I(A\rangle BX)_\rho\Big) \\
&= H(A|X)_\rho + (\mu+1)H(B)_\rho + \lambda H(B|X)_\rho - (\lambda+\mu+1)H(E|X)_\rho && (24.159)\\
&\leq (\mu+1)H(B)_\sigma + H(A|X)_\sigma + \lambda H(B|X)_\sigma - (\lambda+\mu+1)H(E|X)_\sigma && (24.160)\\
&= (\mu+1) + H(A|X)_\sigma + \lambda H(B|X)_\sigma - (\lambda+\mu+1)H(E|X)_\sigma && (24.161)\\
&= (\mu+1) + \sum_x p_X(x)\Big[H(A)_{\phi_x} + \lambda H(B)_{\phi_x} - (\lambda+\mu+1)H(E)_{\phi_x}\Big] && (24.162)\\
&\leq (\mu+1) + \max_x\Big[H(A)_{\phi_x} + \lambda H(B)_{\phi_x} - (\lambda+\mu+1)H(E)_{\phi_x}\Big] && (24.163)\\
&= (\mu+1) + H(A)_{\phi_x^*} + \lambda H(B)_{\phi_x^*} - (\lambda+\mu+1)H(E)_{\phi_x^*}. && (24.164)
\end{aligned}$$

The first equality follows by standard entropic manipulations. The second equality follows because the conditional entropy $H(B|X)$ is invariant under a bit-flipping unitary on the input state that commutes with the channel: $H(B)_{X\rho_x^B X} = H(B)_{\rho_x^B}$. Furthermore, a bit flip on the input state does not change the eigenvalues for the output of the dephasing channel's complementary channel:

$$H(E)_{\mathcal{N}^c(X\rho_x^{A'}X)} = H(E)_{\mathcal{N}^c(\rho_x^{A'})}. \tag{24.165}$$

The first inequality follows because entropy is concave, i.e., the local state σ^B is a mixed version of ρ^B. The third equality follows because

$$H(B)_{\sigma^B} = H\left(\sum_x \frac{1}{2}p_X(x)(\rho_x^B + X\rho_x^B X)\right) = H\left(\frac{1}{2}\sum_x p_X(x)I\right) = 1. \tag{24.166}$$

The fourth equality follows because the system X is classical. The second inequality follows because the maximum value of a realization of a random variable is not less than its expectation. The final equality simply follows by defining ϕ_x^* to be the conditional density

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

operator on systems A, B, and E that arises from sending through the channel a state whose reduction to A' is of the form $\nu|0\rangle\langle 0|^{A'} + (1-\nu)|1\rangle\langle 1|^{A'}$. Thus, an ensemble of the kind in (24.154) is sufficient to attain a point on the boundary of the region.

Evaluating the entropic quantities in Theorem 24.2.1 on a state of the above form then gives the expression for the region in Theorem 24.5.2.

24.5.3 The Lossy Bosonic Channel

One of the most important practical channels in quantum communication is known as the lossy bosonic channel. This channel can model the communication of photons through free space or over a fiber optic cable because the main source of noise in these settings is just the loss of photons. The lossy bosonic channel has one parameter $\eta \in [0,1]$ that characterizes the fraction of photons that make it through the channel to the receiver on average. The environment Eve is able to collect all of the photons that do not make it to the receiver—this fraction is $1-\eta$. Usually, we also restrict the mean number of photons that the sender is allowed to send through the channel (if we do not do so, then there could be an infinite amount of energy available, which is unphysical from a practical perspective, and furthermore, some of the capacities become infinite, which is less interesting from a theoretical perspective). So, we let N_S be the mean number of photons available at the transmitter. Capacities of this channel are then a function of these two parameters η and N_S.

Exercise 24.5.2 Prove that the quantum capacity of a lossy bosonic channel vanishes when $\eta = 1/2$.

In this section, we show how trade-off coding for this channel can give a remarkable gain over time-sharing. Trade-off coding for this channel amounts to a power-sharing strategy, in which the sender dedicates a fraction λ of the available photons to the quantum part of the code and the other fraction $1-\lambda$ to the classical part of the code. This power-sharing strategy is provably optimal (up to a long-standing conjecture) and can beat time-sharing by significant margins (much more so than the dephasing channel does, for example). Specifically, recall that a trade-off coding strategy has the sender and receiver generate random codes from an ensemble of the following form:

$$\left\{p_X(x), |\phi_x\rangle^{AA'}\right\}, \tag{24.167}$$

where $p_X(x)$ is some distribution and the states $|\phi_x\rangle^{AA'}$ are correlated with this distribution, with Alice feeding system A' into the channel. For the lossy bosonic channel, it turns out that the best ensemble to choose is of the following form:

$$\left\{p_{(1-\lambda)N_S}(\alpha), D^{A'}(\alpha)|\psi_{\text{TMS}}\rangle^{AA'}\right\}, \tag{24.168}$$

where α is a complex variable. The distribution $p_{(1-\lambda)N_S}(\alpha)$ is an isotropic Gaussian distribution with variance $(1-\lambda)N_S$:

$$p_{(1-\lambda)N_S}(\alpha) \equiv \frac{1}{\pi(1-\lambda)N_S} \exp\left\{-|\alpha|^2/[(1-\lambda)N_S]\right\}, \tag{24.169}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

where $\lambda \in [0,1]$ is the power-sharing or photon-number-sharing parameter, indicating how many photons to dedicate to the quantum part of the code, while $1-\lambda$ indicates how many photons to dedicate to the classical part. In (24.168), $D^{A'}(\alpha)$ is a "displacement" unitary operator acting on system A' (more on this below), and $|\psi_{\text{TMS}}\rangle^{AA'}$ is a "two-mode squeezed" (TMS) state of the following form:

$$|\psi_{\text{TMS}}\rangle^{AA'} \equiv \sum_{n=0}^{\infty} \sqrt{\frac{[\lambda N_S]^n}{[\lambda N_S + 1]^{n+1}}} |n\rangle^A |n\rangle^{A'}. \tag{24.170}$$

Let θ denote the state resulting from tracing over the mode A:

$$\theta \equiv \text{Tr}_A\left\{ |\psi_{\text{TMS}}\rangle\langle\psi_{\text{TMS}}|^{AA'} \right\} \tag{24.171}$$

$$= \sum_{n=0}^{\infty} \frac{[\lambda N_S]^n}{[\lambda N_S + 1]^{n+1}} |n\rangle\langle n|^{A'}. \tag{24.172}$$

The reduced state θ is known as a thermal state with mean photon number λN_S. We can readily check that its mean photon number is λN_S simply by computing the expectation of the photon number n with respect to the geometric distribution $[\lambda N_S]^n/[\lambda N_S + 1]^{n+1}$:

$$\sum_{n=0}^{\infty} n \frac{[\lambda N_S]^n}{[\lambda N_S + 1]^{n+1}} = \lambda N_S. \tag{24.173}$$

The most important property of the displacement operators $D^{A'}(\alpha)$ for our purposes is that averaging over a random choice of them according to the Gaussian distribution $p_{(1-\lambda)N_S}(\alpha)$, where each operator acts on the state θ, gives a thermal state with mean photon number N_S:

$$\overline{\theta} \equiv \int d\alpha \; p_{(1-\lambda)N_S}(\alpha) \; D(\alpha)\theta D^\dagger(\alpha) \tag{24.174}$$

$$= \sum_{n=0}^{\infty} \frac{[N_S]^n}{[N_S + 1]^{n+1}} |n\rangle\langle n|^{A'}. \tag{24.175}$$

Thus, the choice of ensemble in (24.168) meets the constraint that the average number of photons input to the channel be equal to N_S.

In order to calculate the quantum dynamic capacity region for this lossy bosonic channel, it is helpful to observe that the entropy of a thermal state with mean number of photons N_S is equal to

$$g(N_S) \equiv (N_S + 1)\log_2(N_S + 1) - N_S \log_2(N_S), \tag{24.176}$$

because we will evaluate all of the relevant entropies on thermal states. From Exercise 24.2.1,

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

we know that we should evaluate just the following four entropies:

$$H(A|X)_\sigma = \int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H\big(D(\alpha)\theta D^\dagger(\alpha)\big), \tag{24.177}$$

$$H(B)_\sigma = H\big(\mathcal{N}(\overline{\theta})\big), \tag{24.178}$$

$$H(B|X)_\sigma = \int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H\big(\mathcal{N}(D(\alpha)\theta D^\dagger(\alpha))\big), \tag{24.179}$$

$$H(E|X)_\sigma = \int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H\big(\mathcal{N}^c(D(\alpha)\theta D^\dagger(\alpha))\big), \tag{24.180}$$

where $\mathcal{N}$ is the lossy bosonic channel that transmits η of the input photons to the receiver and $\mathcal{N}^c$ is the complementary channel that transmits $1-\eta$ of the input photons to the environment Eve. We proceed with calculating the above four entropies:

$$\int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H\big(D(\alpha)\theta D^\dagger(\alpha)\big) = \int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H(\theta) \tag{24.181}$$

$$= H(\theta) \tag{24.182}$$

$$= g(\lambda N_S) \tag{24.183}$$

The first equality follows because $D(\alpha)$ is a unitary operator, and the third equality follows because θ is a thermal state with mean photon number N_S. Continuing, we have

$$H\big(\mathcal{N}(\overline{\theta})\big) = g(\eta N_S), \tag{24.184}$$

because $\overline{\theta}$ is a thermal state with mean photon number N_S, but the channel only lets a fraction η of the input photons through on average. The third entropy in (24.179) is equal to

$$\int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H\big(\mathcal{N}(D(\alpha)\theta D^\dagger(\alpha))\big)$$

$$= \int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H\big(D(\alpha)\mathcal{N}(\theta)D^\dagger(\alpha)\big) \tag{24.185}$$

$$= \int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H(\mathcal{N}(\theta)) \tag{24.186}$$

$$= H(\mathcal{N}(\theta)) \tag{24.187}$$

$$= g(\lambda\eta N_S) \tag{24.188}$$

The first equality follows because a displacement operator commutes with the channel (we do not justify this rigorously here). The second equality follows because $D(\alpha)$ is a unitary operator. The final equality follows because θ is a thermal state with mean photon number λN_S, but the channel only lets a fraction η of the inputs photons through on average. By the same line of reasoning (except that the complementary channel lets through only a

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

fraction $1-\eta$ of the input photons), the fourth entropy in (24.180) is equal to

$$\begin{aligned}
&\int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H\big(\mathcal{N}^c(D(\alpha)\theta D^\dagger(\alpha))\big) \\
&= \int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H\big(D(\alpha)\mathcal{N}^c(\theta)D^\dagger(\alpha)\big) && (24.189)\\
&= \int d\alpha\ p_{(1-\lambda)N_S}(\alpha)\ H(\mathcal{N}^c(\theta)) && (24.190)\\
&= H(\mathcal{N}^c(\theta)) && (24.191)\\
&= g(\lambda(1-\eta)N_S). && (24.192)
\end{aligned}$$

Then, by the result of Exercise 24.2.1 and a matching converse that holds whenever $\eta \geq 1/2$,[2] we have the following characterization of the quantum dynamic capacity region of the lossy bosonic channel.

Theorem 24.5.3. *The quantum dynamic capacity region for a lossy bosonic channel with transmissivity $\eta \geq 1/2$ is the union of regions of the form:*

$$C + 2Q \leq g(\lambda N_S) + g(\eta N_S) - g((1-\eta)\lambda N_S), \quad (24.193)$$
$$Q + E \leq g(\eta\lambda N_S) - g((1-\eta)\lambda N_S), \quad (24.194)$$
$$C + Q + E \leq g(\eta N_S) - g((1-\eta)\lambda N_S), \quad (24.195)$$

where $\lambda \in [0,1]$ is a photon-number-sharing parameter and $g(N)$ is the entropy of a thermal state with mean photon number N defined in (24.176). The region is still achievable if $\eta < 1/2$.

Figure 24.5 depicts two important special cases of the region in the above theorem: (a) the trade-off between classical and quantum communication without entanglement assistance and (b) the trade-off between entanglement-assisted and unassisted classical communication. The figure indicates the remarkable improvement over time-sharing that trade-off coding gives.

Other special cases of the above capacity region are the unassisted classical capacity $g(\eta N_S)$ when $\lambda, Q, E = 0$, the quantum capacity $g(\eta N_S) - g((1-\eta)N_S)$ when $\lambda = 1$, $C, E = 0$, and the entanglement-assisted classical capacity $g(N_S) + g(\eta N_S) - g((1-\eta)N_S)$ when $\lambda = 1$, $Q = 0$, and $E = -\infty$.

24.6 History and Further Reading

Shor considered the classical capacity of a channel assisted by a finite amount of shared entanglement [229]. He calculated a trade-off curve that determines how a sender can optimally

[2] We should clarify that the converse holds only if a long-standing minimum-output entropy conjecture is true (researchers have collected much evidence that it should be true).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

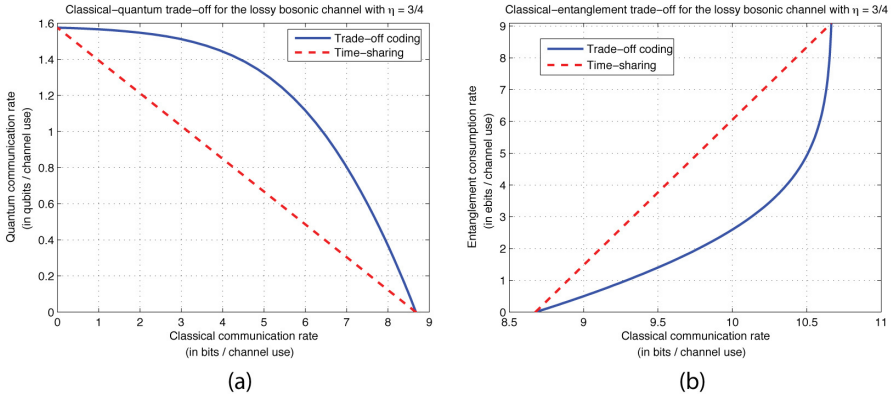

Figure 24.5: (a) Suppose a channel transmits on average 3/4 of the photons to the receiver, while losing the other 1/4 en route. Such a channel can reliably transmit a maximum of $\log_2(3/4) - \log_2(1/4) \approx 1.58$ qubits per channel use, and a mean photon budget of about 200 photons per channel use at the transmitter is sufficient to nearly achieve this quantum capacity. A trade-off coding strategy which lowers the quantum data rate to about 1.4 qubits per channel use while retaining the same mean photon budget allows for a sender to reliably transmit an additional 4.5 classical bits per channel use, while time-sharing would only allow for an additional 1 classical bit per channel use with this photon budget. The 6.5 dB increase in the classical data rate that trade-off coding gives over time-sharing for this example is strong enough to demand that quantum communication engineers employ trade-off coding strategies in order to take advantage of such theoretical performance gains. (b) The sender and the receiver share entanglement, and the sender would like to transmit classical information while minimizing the consumption of entanglement. With a mean photon budget of 200 photons per channel use over a channel that propagates only 3/4 of the photons input to it, the sender can reliably transmit a maximum of about 10.7 classical bits per channel use while consuming entanglement at a rate of about 9.1 entangled bits per channel use. With trade-off coding, the sender can significantly reduce the entanglement consumption rate to about 5 entangled bits per channel use while still transmitting about 10.5 classical bits per channel use, only a 0.08 dB decrease in the rate of classical communication for a 2.6 dB decrease in the entanglement consumption rate. The savings in entanglement consumption could be useful for them if they would like to have the extra entanglement for future rounds of assisted communication.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

trade the consumption of noiseless entanglement with the generation of noiseless classical communication. This trade-off curve also bounds a rate region consisting of rates of entanglement consumption and generated classical communication. Shor's result then inspired Devetak and Shor to consider a scenario where a sender exploits a noisy quantum channel to simultaneously transmit both noiseless classical and quantum information [73], a scenario later dubbed "classically-enhanced quantum coding" [159, 160] after schemes formulated in the theory of quantum error correction [178, 249]. Devetak and Shor provided a multi-letter characterization of the classically-enhanced quantum capacity region for general channels, but they were able to show that both generalized dephasing channels and erasure channels admit single-letter capacity regions.

The above scenarios are a part of the dynamic, double-resource quantum Shannon theory, where a sender can exploit a noisy quantum channel to generate two noiseless resources, or a sender can exploit a noisy quantum channel in addition to a noiseless resource to generate another noiseless resource. This theory culminated with the work of Devetak *et al.* that provided a multi-letter characterization for virtually every permutation of two resources and a noisy quantum channel which one can consider [70, 71]. Other researchers concurrently considered how noiseless resources might trade off against each other in tasks outside of the dynamic, double-resource quantum Shannon theory, such as quantum compression [14, 133, 176], remote state preparation [32, 4], and hybrid quantum memories [179].

Refs. [159, 160, 253] considered the dynamic, triple-resource quantum Shannon theory by providing a multi-letter characterization of an entanglement-assisted quantum channel's ability to transmit both classical and quantum information. Ref. [159] also constructed a new protocol, dubbed the "classically-enhanced father protocol," that outperforms a time-sharing strategy for transmitting both classical and quantum information over an entanglement-assisted quantum channel. Brádler *et al.* showed that the quantum Hadamard channels have a single-letter capacity region [46]. Later studies continued these efforts of exploring information trade-offs [164, 252].

Ref. [250] recently found the quantum dynamic capacity region of the lossy bosonic channel (up to a long-standing minimum-output entropy conjecture). The results there build on a tremendous body of literature for bosonic channels. Giovannetti *et al.* found the classical capacity of the lossy bosonic channel [104]. Others found the entanglement-assisted classical and quantum capacities of the lossy bosonic channel [34, 141, 107, 106] and its quantum capacity [263, 120]. The long-standing minimum-output entropy conjecture is detailed in Refs. [103, 120, 105].

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

CHAPTER 25

Summary and Outlook

This brief final chapter serves as a compact summary of all the results presented in this book, it highlights information processing tasks that we did not cover, and it discusses new directions. We exploit the resource inequality formalism in our summary.

A resource inequality is a statement of achievability:

$$\sum_k \alpha_k \geq \sum_j \beta_j, \tag{25.1}$$

meaning that the resources $\{\alpha_k\}$ on the LHS can simulate the resources $\{\beta_j\}$ on the RHS. The simulation can be exact and finite or asymptotically perfect. We can classify resources as follows:

1. Unit, noiseless, or noisy.

2. Dynamic or static. Moreover, dynamic resources can be *relative* (see below).

3. Classical, quantum, or hybrid.

The unit resources are as follows: $[c \to c]$ represents one noiseless classical bit channel, $[q \to q]$ represents one noiseless qubit channel, $[qq]$ represents one noiseless ebit, and $[q \to qq]$ represents one noiseless coherent bit channel. We also have $[c \to c]_{\text{priv}}$ as a noiseless private classical bit channel and $[cc]_{\text{priv}}$ as a noiseless bit of secret key. An example of a noiseless resource is a pure bipartite state $|\phi\rangle^{AB}$ shared between Alice and Bob or an identity channel $I^{A\to B}$ from Alice to Bob. An example of a noisy resource could be a mixed bipartite state ρ^{AB} or a noisy channel $\mathcal{N}^{A'\to B}$. Unit resources are a special case of noiseless resources, which are in turn a special case of noisy resources.

A shared state ρ^{AB} is an example of a noisy static resource, and a channel $\mathcal{N}$ is an example of a noisy dynamic resource. We indicate these by $\langle\rho\rangle$ or $\langle\mathcal{N}\rangle$ in a resource inequality. We can be more precise if necessary and write $\langle\mathcal{N}\rangle$ as a dynamic, relative resource $\langle\mathcal{N}^{A'\to B} : \sigma^{A'}\rangle$, meaning that the protocol only works as it should if the state input to the channel is $\sigma^{A'}$.

It is obvious when a resource is classical or when it is quantum, and an example of a hybrid resource is a classical-quantum state

$$\rho^{XA} = \sum_x p_X(x)|x\rangle\langle x|^X \otimes \rho_x^A. \tag{25.2}$$

25.1 Unit Protocols

Chapter 6 discussed entanglement distribution

$$[q \to q] \geq [qq], \tag{25.3}$$

teleportation

$$2[c \to c] + [qq] \geq [q \to q], \tag{25.4}$$

and super-dense coding

$$[q \to q] + [qq] \geq 2[c \to c]. \tag{25.5}$$

Chapter 7 introduced coherent dense coding

$$[q \to q] + [qq] \geq 2[q \to qq], \tag{25.6}$$

and coherent teleportation

$$2[q \to qq] \geq [q \to q] + [qq]. \tag{25.7}$$

The fact that these two resource inequalities are dual under resource reversal implies the important coherent communication identity:

$$2[q \to qq] = [q \to q] + [qq]. \tag{25.8}$$

We also have the following resource inequalities:

$$[q \to q] \geq [q \to qq] \geq [qq]. \tag{25.9}$$

Other unit protocols not covered in this book are the one-time pad

$$[c \to c]_{\text{pub}} + [cc]_{\text{priv}} \geq [c \to c]_{\text{priv}}, \tag{25.10}$$

secret key distribution

$$[c \to c]_{\text{priv}} \geq [cc]_{\text{priv}}, \tag{25.11}$$

and private-to-public transmission

$$[c \to c]_{\text{priv}} \geq [c \to c]_{\text{pub}}. \tag{25.12}$$

The last protocol assumes a model where the receiver can locally copy information and place it in a register to which Eve has access.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

25.2 Noiseless Quantum Shannon Theory

Noiseless quantum Shannon theory consists of resource inequalities involving unit resources and one non-unit, noiseless resource, such as an identity channel or a pure bipartite state.

Schumacher compression from Chapter 17 gives a way to simulate an identity channel $I^{A\to B}$ acting on a mixed state ρ^A by exploiting noiseless qubit channels at a rate equal to the entropy $H(A)_\rho$:

$$H(A)_\rho[q \to q] \geq \left\langle I^{A\to B} : \rho^A \right\rangle. \tag{25.13}$$

We also know that if n uses of an identity channel are available, then achievability of the coherent information for quantum communication (Chapter 23) implies that we can send quantum data down this channel at a rate equal to $H(B) - H(E)$, where the entropies are with respect to some input density operator ρ^A. But $H(E) = 0$ because the channel is the identity channel (the environment gets no information) and $H(B) = H(A)_\rho$ because Alice's input goes directly to Bob. This gives us the following resource inequality:

$$\left\langle I^{A\to B} : \rho^A \right\rangle \geq H(A)_\rho[q \to q], \tag{25.14}$$

and combining (25.13) and (25.14) gives the following resource equality:

$$\left\langle I^{A\to B} : \rho^A \right\rangle = H(A)_\rho[q \to q]. \tag{25.15}$$

Entanglement concentration from Chapter 18 converts many copies of a pure, bipartite state $|\phi\rangle^{AB}$ into ebits at a rate equal to the entropy of entanglement:

$$\left\langle \phi^{AB} \right\rangle \geq H(A)_\phi[qq]. \tag{25.16}$$

We did not discuss entanglement dilution in any detail [188, 187, 124, 135], but it is a protocol that exploits a sublinear amount of classical communication to dilute ebits into n copies of a pure, bipartite state $|\phi\rangle^{AB}$. Ignoring the sublinear rate of classical communication gives the following resource inequality:

$$H(A)_\phi[qq] \geq \left\langle \phi^{AB} \right\rangle. \tag{25.17}$$

Combining entanglement concentration and entanglement dilution gives the following resource equality:

$$\left\langle \phi^{AB} \right\rangle = H(A)_\phi[qq]. \tag{25.18}$$

The noiseless quantum Shannon theory is satisfactory in the sense that we can obtain resource *equalities*, illustrating the interconvertibility of noiseless qubit channels with a relative identity channel and pure, bipartite states with ebits.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

25.3 Noisy Quantum Shannon Theory

Noisy quantum Shannon theory has resource inequalities with one noisy resource, such as a noisy channel or a noisy state, interacting with other unit resources. We can further classify a resource inequality as dynamic or static, depending on whether the noisy resource involved is dynamic or static.

We first review the dynamic resource inequalities presented in this book. These protocols involve a noisy channel interacting with the other unit resources. Many of the protocols in noisy quantum Shannon theory generate random codes from a state of the following form:

$$\rho^{XABE} \equiv \sum_x p_X(x)|x\rangle\langle x|^X \otimes U_{\mathcal{N}}^{A'\to BE}(\phi_x^{AA'}), \tag{25.19}$$

where $\phi_x^{AA'}$ is a pure, bipartite state and $U_{\mathcal{N}}^{A'\to BE}$ is an isometric extension of a channel $\mathcal{N}^{A'\to B}$. Also important is a special case of the above form:

$$\sigma^{ABE} \equiv U_{\mathcal{N}}^{A'\to BE}(\phi^{AA'}), \tag{25.20}$$

where $\phi^{AA'}$ is a pure, bipartite state. Holevo-Schumacher-Westmoreland coding for classical communication over a quantum channel (Chapter 19) is the following resource inequality:

$$\langle\mathcal{N}\rangle \geq I(X;B)_\rho[c\to c]. \tag{25.21}$$

Devetak-Cai-Winter-Yeung coding for private classical communication over a quantum channel (Chapter 22) is as follows:

$$\langle\mathcal{N}\rangle \geq \Big(I(X;B)_\rho - I(X;E)_\rho\Big)[c\to c]_{\text{priv}}. \tag{25.22}$$

Upgrading the private classical code to one that operates coherently gives Devetak's method for coherent communication over a quantum channel (Chapter 23):

$$\langle\mathcal{N}\rangle \geq I(A\rangle B)_\sigma[q\to qq], \tag{25.23}$$

which we showed can be converted asymptotically into a protocol for quantum communication:

$$\langle\mathcal{N}\rangle \geq I(A\rangle B)_\sigma[q\to q]. \tag{25.24}$$

Bennett-Shor-Smolin-Thapliyal coding for entanglement-assisted classical communication over a quantum channel (Chapter 20) is the following resource inequality:

$$\langle\mathcal{N}\rangle + H(A)_\sigma[qq] \geq I(A;B)_\sigma[c\to c]. \tag{25.25}$$

We showed how to upgrade this protocol to one for entanglement-assisted coherent communication (Chapter 21):

$$\langle\mathcal{N}\rangle + H(A)_\sigma[qq] \geq I(A;B)_\sigma[q\to qq], \tag{25.26}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

and combining with the coherent communication identity gives the following protocol for entanglement-assisted quantum communication:

$$\langle \mathcal{N} \rangle + \frac{1}{2} I(A;E)_\sigma [qq] \geq \frac{1}{2} I(A;B)_\sigma [q \rightarrow q]. \tag{25.27}$$

Further combining with entanglement distribution gives the resource inequality in (25.24) for quantum communication. By combining the HSW and BSST protocols together (this needs to be done at the level of coding and not at the level of resource inequalities—see Chapter 21), we recover a protocol for entanglement-assisted communication of classical and quantum information:

$$\langle \mathcal{N} \rangle + \frac{1}{2} I(A;E|X)_\sigma [qq] \geq \frac{1}{2} I(A;B|X)_\sigma [q \rightarrow q] + I(X;B)_\sigma [c \rightarrow c]. \tag{25.28}$$

This protocol recovers any protocol in dynamic quantum Shannon theory that involves a noisy channel and the three unit resources after combining it with the three unit protocols in (25.3-25.5). Important special cases are entanglement-assisted classical communication with limited entanglement:

$$\langle \mathcal{N} \rangle + H(A|X)_\sigma [qq] \geq I(AX;B)_\sigma [c \rightarrow c], \tag{25.29}$$

and simultaneous classical and quantum communication:

$$\langle \mathcal{N} \rangle \geq I(X;B)_\sigma [c \rightarrow c] + I(A\rangle BX)_\sigma [q \rightarrow q]. \tag{25.30}$$

Chapter 21 touched on some important protocols in static quantum Shannon theory. These protocols involve some noisy state ρ^{AB} interacting with the unit resources. The protocol for coherent-assisted state transfer is the static couterpart to the protocol in (25.26):

$$\langle W^{S \rightarrow AB} : \rho^S \rangle + H(A)_\rho [q \rightarrow q] \geq I(A;B)_\rho [q \rightarrow qq] + \langle I^{S \rightarrow \hat{B}B} : \rho^S \rangle, \tag{25.31}$$

where W is some isometry that distributes the state from a source S to two parties A and B and $I^{S \rightarrow \hat{B}B}$ is the identity. Ignoring the source and state transfer in the above protocol gives a protocol for quantum-assisted coherent communication:

$$\langle \rho \rangle + H(A)_\rho [q \rightarrow q] \geq I(A;B)_\rho [q \rightarrow qq]. \tag{25.32}$$

We can also combine (25.31) with the unit protocols to obtain quantum-assisted state transfer:

$$\langle W^{S \rightarrow AB} : \rho^S \rangle + \frac{1}{2} I(A;R)_\varphi [q \rightarrow q] \geq \frac{1}{2} I(A;B)_\varphi [qq] + \langle I^{S \rightarrow \hat{B}B} : \rho^S \rangle, \tag{25.33}$$

and classical-assisted state transfer:

$$\langle W^{S \rightarrow AB} : \rho^S \rangle + I(A;R)_\varphi [c \rightarrow c] \geq I(A\rangle B)_\varphi [qq] + \langle I^{S \rightarrow \hat{B}B} : \rho^S \rangle, \tag{25.34}$$

where $|\varphi\rangle^{ABR}$ is a purification of ρ^{AB}. We also have noisy super-dense coding

$$\langle \rho \rangle + H(A)_\rho [q \rightarrow q] \geq I(A;B)_\rho [c \rightarrow c], \tag{25.35}$$

and noisy teleportation

$$\langle \rho \rangle + I(A;B)_\rho [c \rightarrow c] \geq I(A\rangle B)_\rho [q \rightarrow q], \tag{25.36}$$

by combining (25.32) with the coherent communication identity and the unit protocols.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

25.4 Protocols not covered in this book

There are many important protocols that we did not cover in this book because our focus here was on channels. One such example is *quantum state redistribution*. Suppose that Alice and Bob share many copies of a tripartite state ρ^{ACB} where Alice has the shares AC and Bob has the share B. The goal of state redistribution is for Alice to transfer the C part of the state to Bob using the minimal resources needed to do so. It is useful to identify a pure state φ^{RACB} as a purification of ρ^{ACB}, where R is the purifying system. Devetak and Yard showed the existence of the following state redistribution protocol [77, 267]:

$$\langle W^{S\to AC|B} : \rho^S \rangle + \frac{1}{2} I(C;RB)_\varphi[q \to q] + \frac{1}{2} I(C;A)_\varphi[qq] \geq \langle W^{S\to A|CB} : \rho^S \rangle + \frac{1}{2} I(C;B)_\varphi[q \to q] + \frac{1}{2} I(C;B)_\varphi[qq], \quad (25.37)$$

where $W^{S\to AC|B}$ is some isometry that distributes the system S as AC for Alice and B for Bob and $W^{S\to A|CB}$ is defined similarly. They also demonstrated that the above resource inequality gives an optimal cost pair for the quantum communication rate Q and the entanglement consumption rate E, with

$$Q = \frac{1}{2} I(C;R|B)_\varphi, \quad (25.38)$$

$$E = \frac{1}{2}\Big[I(C;A)_\varphi - I(C;B)_\varphi \Big]. \quad (25.39)$$

Thus, their protocol gives a direct operational interpretation to the conditional quantum mutual information $\frac{1}{2} I(C;R|B)_\varphi$ as the net rate of quantum communication required in quantum state redistribution.

A simple version of the quantum reverse Shannon theorem gives a way to simulate the action of a channel $\mathcal{N}^{A'\to B}$ on some input state $\rho^{A'}$ by exploiting classical communication and entanglement [34, 27, 39]:

$$H(B)_\sigma[qq] + I(R;B)_\sigma[c \to c] \geq \langle \mathcal{N}^{A'\to B} : \rho^{A'} \rangle, \quad (25.40)$$

where

$$\sigma^{RB} \equiv \mathcal{N}^{A'\to B}(\varphi^{RA'}), \quad (25.41)$$

with $\varphi^{RA'}$ a purification of $\rho^{A'}$. One utility of the quantum reverse Shannon theorem is that it gives an indication of how one channel might simulate another in the presence of shared entanglement. In the simulation of the channel $\mathcal{N}^{A'\to B}$, the environment is also simulated and ends up in Alice's possession. Thus, they end up simulating the quantum feedback channel $U_{\mathcal{N}}^{A'\to AB}$, and we can restate (25.40) as follows:

$$H(B)_\sigma[qq] + I(R;B)_\sigma[c \to c] \geq \langle U_{\mathcal{N}}^{A'\to AB} : \rho^{A'} \rangle. \quad (25.42)$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

It is possible to upgrade the classical communication to coherent communication [69], leading to the following coherent, fully-quantum version of the quantum reverse Shannon theorem [3]:

$$\frac{1}{2}I(A;B)_\sigma[qq] + \frac{1}{2}I(R;B)_\sigma[q \to q] \geq \langle U_{\mathcal{N}}^{A' \to AB} : \rho^{A'} \rangle. \tag{25.43}$$

Combining this resource inequality with the following one from Exercise 21.1.1

$$\langle U_{\mathcal{N}}^{A' \to AB} : \rho^{A'} \rangle \geq \frac{1}{2}I(A;B)_\sigma[qq] + \frac{1}{2}I(R;B)_\sigma[q \to q] \tag{25.44}$$

gives the following satsifying resource equality:

$$\langle U_{\mathcal{N}}^{A' \to AB} : \rho^{A'} \rangle = \frac{1}{2}I(A;B)_\sigma[qq] + \frac{1}{2}I(R;B)_\sigma[q \to q]. \tag{25.45}$$

The above resource equality is a generalization of the coherent communication identity. A more general version of the quantum reverse Shannon theorem quantifies the resources needed to simulate many independent instances of a quantum channel on an arbitrary input state, and the proof in this case is significantly more complicated [27, 39].

Other protocols that we did not cover are remote state preparation [28, 32, 4], classical compression with quantum side information [74], trade-offs between public and private resources and channels [252], trade-offs in compression [133], and a trade-off for a noisy state with the three unit resources [160]. The resource inequality formalism is helpful for devising new protocols in quantum Shannon theory by imagining some resources to be unit and others to be noisy.

25.5 Network Quantum Shannon Theory

The field of network quantum Shannon theory has arisen in recent years, motivated by the idea that one day we will be dealing with a quantum Internet in which channels of increasing complexity can connect a number of senders to a number of receivers. A quantum multiple access channel has multiple senders and one receiver. Various authors have considered classical communication over a multiple access channel [257], quantum communication over multiple access channels [148, 269], and entanglement-assisted protocols [156]. A quantum broadcast channel has one sender and multiple receivers. Various authors have addressed similar scenarios in this setting [270, 84]. A quantum interference channel has multiple senders and multiple receivers in which certain sender-receiver pairs are interested in communicating. Recent progress in this direction is in Ref. [93]. One could also consider distributed compression tasks, and various authors have contributed to this direction [8, 3, 210]. We could imagine a future textbook containing several chapters that summarize all of the progress in network quantum Shannon theory and the novel techniques needed to handle coding over such channels. Savov highlights much of this direction in his PhD thesis [211] (at least for classical communication).

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

25.6 Future Directions

Quantum Shannon theory has evolved from the first and simplest result regarding Schumacher compression to a whole host of protocols that indicate how much data we can transmit over noisy quantum channels or how much we can compress information of varying types—the central question in any task is, "How many unit resources can we extract from a given non-unit resource, perhaps with the help of other non-unit resources?" This book may give the impression that so much has been solved in the area of quantum Shannon theory that little remains for the future, but this is actually far from the truth! There remains much to do to improve our understanding, and this final section briefly outlines just a few of these important questions.

Find a better formula for the classical capacity other than the HSW formula. Our best characterization of the classical capacity is with a regularized version of the HSW formula, and this is unsatisfying in several ways that we have mentioned before. In a similar vein, find a better formula for the private classical capacity, the quantum capacity, and even for the trade-off capacities. All of these formulas are unsatisfying because their regularizations seem to be necessary in the general case. It could be the case that an entropic expression evaluated on some finite tensor power of the channels would be sufficient to characterize the capacity for different tasks, but this is a difficult question to answer. Interestingly, recent work suggests pursuing to find out whether this question is algorithmically undecidable (see Ref. [261]). Effects such as superactivation of quantum capacity (see Section 23.7.2) and non-additivity of private capacity (see Section 22.5.2) have highlighted how little we actually know about the corresponding information processing tasks in the general case. Also, it is important to understand these effects more fully and to see if there is any way of exploiting them in a practical communication scheme. Finally, a different direction is to expand the number of channels that have additive capacities. For example, finding the quantum capacity of a non-degradable quantum channel would be a great result.

Continue to explore network quantum Shannon theory. The single-sender, single-receiver channel setting is a useful model for study and applies to many practical scenarios, but eventually, we will be dealing with channels connecting many inputs to many outputs. Having such an understanding for information transmission in these scenarios could help guide the design of practical communication schemes and might even shed light on the open problems in the preceding paragraph.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

APPENDIX A

Miscellaneous Mathematics

This section collects various useful definitions and lemmas that we use throughout the proofs of certain theorems in this book.

Lemma A.0.1. *Suppose that M and N are positive operators. Then the operators $M+N$, MNM, and NMN are positive.*

Lemma A.0.2. *Suppose that the operators $\hat{\omega}$ and ω have trace less than or equal to unity. Suppose $\hat{\omega}$ lies in the operator interval $[(1-\epsilon)\omega, (1+\epsilon)\omega]$. Then*

$$\|\hat{\omega} - \omega\|_1 \leq \epsilon. \tag{A.1}$$

Proof. The statement "$\hat{\omega}$ lies in the operator interval $[(1-\epsilon)\omega, (1+\epsilon)\omega]$" is equivalent to the following two conditions:

$$(1+\epsilon)\omega - \hat{\omega} = \epsilon\omega - (\hat{\omega} - \omega) \geq 0, \tag{A.2}$$
$$\hat{\omega} - (1-\epsilon)\omega = (\hat{\omega} - \omega) + \epsilon\omega \geq 0. \tag{A.3}$$

Let $\alpha \equiv \hat{\omega} - \omega$. Let us rewrite α in terms of the positive operators α^+ and α^-

$$\alpha = \alpha^+ - \alpha^- \tag{A.4}$$

as we did in the proof of Lemma 9.1.1. The above conditions become as follows:

$$\epsilon\omega - \alpha \geq 0, \tag{A.5}$$
$$\alpha + \epsilon\omega \geq 0. \tag{A.6}$$

Let the positive projectors Π^+ and Π^- project onto the respective supports of α^+ and α^-. We then apply the projector Π^+ to the first condition:

$$\Pi^+(\epsilon\omega - \alpha)\Pi^+ \geq 0 \tag{A.7}$$
$$\Rightarrow \epsilon\Pi^+\omega\Pi^+ - \Pi^+\alpha\Pi^+ \geq 0 \tag{A.8}$$
$$\Rightarrow \epsilon\Pi^+\omega\Pi^+ - \alpha^+ \geq 0 \tag{A.9}$$

where the first inequality follows from Lemma A.0.1. We apply the projector Π^- to the second condition:

$$\Pi^-(\alpha + \epsilon\omega)\Pi^- \geq 0 \tag{A.10}$$
$$\Rightarrow \Pi^-\alpha\Pi^- + \epsilon\Pi^-\omega\Pi^- \geq 0 \tag{A.11}$$
$$\Rightarrow -\alpha^- + \epsilon\Pi^-\omega\Pi^- \geq 0 \tag{A.12}$$

where the first inequality again follows from Lemma A.0.1. Adding the two positive operators together gives another positive operator by Lemma A.0.1:

$$\epsilon\Pi^+\omega\Pi^+ - \alpha^+ - \alpha^- + \epsilon\Pi^-\omega\Pi^- \geq 0 \tag{A.13}$$
$$\Rightarrow \epsilon\Pi^+\omega\Pi^+ - |\hat{\omega} - \omega| + \epsilon\Pi^-\omega\Pi^- \geq 0 \tag{A.14}$$
$$\Rightarrow \epsilon\omega - |\hat{\omega} - \omega| \geq 0 \tag{A.15}$$

Apply the trace operation to get the following inequality:

$$\epsilon\text{Tr}\{\omega\} \geq \text{Tr}\{|\hat{\omega} - \omega|\} = \|\hat{\omega} - \omega\|_1 \tag{A.16}$$

Using the hypothesis that $\text{Tr}\{\omega\} \leq 1$ gives the desired result. □

Theorem A.0.1 (Polar Decomposition)**.** *Any operator A admits a left polar decomposition:*

$$A = U\sqrt{A^\dagger A}, \tag{A.17}$$

and a right polar decomposition:

$$A = \sqrt{AA^\dagger}V. \tag{A.18}$$

Proof. We give a simple proof for just the right polar decomposition by appealing to the singular value decomposition. Any operator A admits a singular value decomposition:

$$A = U_1\Sigma U_2, \tag{A.19}$$

where U_1 and U_2 are unitary operators and Σ is an operator with positive singular values. Then

$$AA^\dagger = U_1\Sigma U_2 U_2^\dagger \Sigma U_1^\dagger = U_1\Sigma^2 U_1^\dagger, \tag{A.20}$$

and thus

$$\sqrt{AA^\dagger} = U_1\Sigma U_1^\dagger. \tag{A.21}$$

We can take $V = U_1U_2$ and we obtain the right polar decomposition of A

$$\sqrt{AA^\dagger}V = U_1\Sigma U_1^\dagger U_1 U_2 = U_1\Sigma U_2 = A. \tag{A.22}$$

□

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Lemma A.0.3. *Let A be any operator and U be a unitary operator. Then*

$$|Tr\{AU\}| \leq Tr\{|A|\}, \tag{A.23}$$

with saturation of the equality when $U = V^\dagger$, where $A = |A|V$ is the right polar decomposition of A.

Proof. It is straightforward to show equality under the scenario stated in the theorem. It holds that

$$|\mathrm{Tr}\{AU\}| = |\mathrm{Tr}\{|A|VU\}| = \left|\mathrm{Tr}\left\{|A|^{\frac{1}{2}}|A|^{\frac{1}{2}}VU\right\}\right|. \tag{A.24}$$

It follows from the Cauchy-Schwarz inequality for the Hilbert-Schmidt inner product that

$$|\mathrm{Tr}\{AU\}| \leq \sqrt{\mathrm{Tr}\{|A|\}\mathrm{Tr}\{U^\dagger V^\dagger |A| VU\}} = \mathrm{Tr}\{|A|\}. \tag{A.25}$$

□

Lemma A.0.4. *Consider two collections of orthonormal states $(|\chi_j\rangle)_{j\in[N]}$ and $(|\zeta_j\rangle)_{j\in[N]}$ such that $\langle\chi_j|\zeta_j\rangle \geq 1-\epsilon$ for all j. There exist phases γ_j and δ_j such that*

$$\langle\hat{\chi}|\hat{\zeta}\rangle \geq 1-\epsilon, \tag{A.26}$$

where

$$|\hat{\chi}\rangle = \frac{1}{\sqrt{N}}\sum_{j=1}^{N} e^{i\gamma_j}|\chi_j\rangle, \tag{A.27}$$

$$|\hat{\zeta}\rangle = \frac{1}{\sqrt{N}}\sum_{j=1}^{N} e^{i\delta_j}|\zeta_j\rangle. \tag{A.28}$$

Proof. Define the Fourier transformed states

$$|\hat{\chi}_s\rangle \equiv \frac{1}{\sqrt{N}}\sum_{j=1}^{N} e^{2\pi ijs/N}|\chi_j\rangle, \tag{A.29}$$

and similarly define $|\hat{\zeta}_s\rangle$. By Parseval's relation, it follows that

$$\frac{1}{N}\sum_{s=1}^{N}\langle\hat{\chi}_s|\hat{\zeta}_s\rangle = \frac{1}{N}\sum_{j=1}^{N}\langle\chi_j|\zeta_j\rangle \geq 1-\epsilon. \tag{A.30}$$

Thus, at least one value of s obeys the following inequality:

$$e^{i\theta_s}\langle\hat{\chi}_s|\hat{\zeta}_s\rangle \geq 1-\epsilon, \tag{A.31}$$

for some phase θ_s. Setting $\gamma_j = 2\pi js/N$ and $\delta_j = \gamma_j + \theta_s$ satisfies the statement of the lemma. □

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

A.1 The Operator Chernoff Bound

In this section, we provide the proof of Ahlswede and Winter's Operator Chernoff Bound from Ref. [7]. Recall that we write $A \geq B$ if $A - B$ is a positive semidefinite operator and we write $A \not\geq B$ otherwise.

Lemma A.1.1 (Operator Chernoff Bound). *Let $\xi_1, \ldots, \xi_M$ be M independent and identically distributed random variables with values in the algebra $\mathcal{B}(\mathcal{H})$ of bounded linear operators on some Hilbert space $\mathcal{H}$. Each ξ_m has all of its eigenvalues between the null operator 0 and the identity operator I:*

$$\forall m \in [M] : 0 \leq \xi_m \leq I. \tag{A.32}$$

Let $\overline{\xi}$ denote the sample average of the M random variables:

$$\overline{\xi} = \frac{1}{M} \sum_{m=1}^{M} \xi_m. \tag{A.33}$$

Suppose that the expectation $\mathbb{E}_\xi\{\xi_m\} \equiv \mu$ of each operator ξ_m exceeds the identity operator scaled by a number $a \in (0, 1)$:

$$\mu \geq aI. \tag{A.34}$$

Then for every η where $0 < \eta < 1/2$ and $(1 + \eta)a \leq 1$, we can bound the probability that the sample average $\overline{\xi}$ lies inside the operator interval $[(1 \pm \eta)\mu]$:

$$\Pr_{\xi}\big\{(1 - \eta)\mu \leq \overline{\xi} \leq (1 + \eta)\mu\big\} \geq 1 - 2 \dim \mathcal{H} \exp\left(-\frac{M\eta^2 a}{4 \ln 2}\right). \tag{A.35}$$

Thus it is highly likely that the sample average operator $\overline{\xi}$ becomes close to the true expected operator μ as M becomes large.

We prove the above lemma in the same way that Ahlswede and Winter did, by making a progression through the Operator Markov inequality all the way to the proof of the above Operator Chernoff Bound.

Lemma A.1.2 (Operator Markov Inequality). *Let X be a random variable with values in the algebra $\mathcal{B}^+(\mathcal{H})$ of positive bounded linear operators on some Hilbert space $\mathcal{H}$. Let $\mathbb{E}\{X\}$ denote its expectation. Let A be a fixed positive operator in $\mathcal{B}^+(\mathcal{H})$. Then*

$$\Pr\{X \not\leq A\} \leq Tr\big\{\mathbb{E}\{X\}A^{-1}\big\}. \tag{A.36}$$

Proof. Observe that if $X \not\leq A$ then $A^{-1/2}XA^{-1/2} \not\leq I$. This then implies that the largest eigenvalue of $A^{-1/2}XA^{-1/2}$ exceeds one: $\left\|A^{-1/2}XA^{-1/2}\right\| > 1$. Let $I_{X \not\leq A}$ denote an indicator function for the event $X \not\leq A$. We then have that

$$I_{X \not\leq A} \leq \mathrm{Tr}\big\{A^{-1/2}XA^{-1/2}\big\}. \tag{A.37}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above inequality follows because the RHS is non-negative if the indicator is zero. If the indicator is one, then the RHS exceeds one because its largest eigenvalue is greater than one and the trace exceeds the largest eigenvalue for a positive operator. We then have the following inequalities:

$$\Pr\{X \not\leq A\} = \mathbb{E}\Big\{I_{X \not\leq A}\Big\} \leq \mathbb{E}\big\{\mathrm{Tr}\big\{A^{-1/2} X A^{-1/2}\big\}\big\} \tag{A.38}$$

$$= \mathbb{E}\big\{\mathrm{Tr}\big\{X A^{-1}\big\}\big\} = \mathrm{Tr}\big\{\mathbb{E}\{X\} A^{-1}\big\}, \tag{A.39}$$

where the first inequality follows from (A.37) and the second equality from cyclicity of trace. □

Lemma A.1.3 (Bernstein Trick)**.** *Let X, X_1, ..., X_n be IID Hermitian random variables in $\mathcal{B}(\mathcal{H})$, and let A be a fixed Hermitian operator. Then for any invertible operator T, we have*

$$\Pr\left\{\sum_{k=1}^{n} X_k \not\leq nA\right\} \leq \dim(\mathcal{H}) \left\| \mathbb{E}\big\{\exp\big\{T(X - A)T^\dagger\big\}\big\} \right\|^n. \tag{A.40}$$

Proof. The proof of this lemma relies on the Golden-Thompson inequality from statistical mechanics which holds for any two Hermitian operators A and B (which we state without proof):

$$\mathrm{Tr}\{\exp\{A + B\}\} \leq \mathrm{Tr}\{\exp\{A\} \exp\{B\}\}. \tag{A.41}$$

Consider the following chain of inequalities:

$$\Pr\left\{\sum_{k=1}^{n} X_k \not\leq nA\right\} = \Pr\left\{\sum_{k=1}^{n} (X_k - A) \not\leq 0\right\} \tag{A.42}$$

$$= \Pr\left\{\sum_{k=1}^{n} T(X_k - A)T^\dagger \not\leq 0\right\} \tag{A.43}$$

$$= \Pr\left\{\exp\left\{\sum_{k=1}^{n} T(X_k - A)T^\dagger\right\} \not\leq I\right\} \tag{A.44}$$

$$\leq \mathrm{Tr}\left\{\mathbb{E}\left\{\exp\left\{\sum_{k=1}^{n} T(X_k - A)T^\dagger\right\}\right\}\right\} \tag{A.45}$$

The first two equalities are straightforward and the third follows because $A \leq B$ is equivalent to $\exp\{A\} \leq \exp\{B\}$ for commuting operators A and B. The first inequality follows from

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

applying the Operator Markov inequality. Continuing, we have

$$= \mathbb{E}\left\{ \operatorname{Tr}\left\{ \exp\left\{ \sum_{k=1}^{n} T(X_k - A)T^\dagger \right\} \right\} \right\} \tag{A.46}$$

$$\leq \mathbb{E}\left\{ \operatorname{Tr}\left\{ \exp\left\{ \sum_{k=1}^{n-1} T(X_k - A)T^\dagger \right\} \exp\{T(X_n - A)T^\dagger\} \right\} \right\} \tag{A.47}$$

$$= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left\{ \operatorname{Tr}\left\{ \exp\left\{ \sum_{k=1}^{n-1} T(X_k - A)T^\dagger \right\} \mathbb{E}_{X_n}\left\{ \exp\{T(X_n - A)T^\dagger\} \right\} \right\} \right\} \tag{A.48}$$

$$= \mathbb{E}_{X_1,\ldots,X_{n-1}}\left\{ \operatorname{Tr}\left\{ \exp\left\{ \sum_{k=1}^{n-1} T(X_k - A)T^\dagger \right\} \mathbb{E}_{X}\left\{ \exp\{T(X - A)T^\dagger\} \right\} \right\} \right\} \tag{A.49}$$

The first equality follows from exchanging the expectation and the trace. The first inequality follows from applying the Golden-Thompson inequality. The second and third equalities follow from the IID assumption. Continuing,

$$\leq \mathbb{E}_{X_1,\ldots,X_{n-1}}\left\{ \operatorname{Tr}\left\{ \exp\left\{ \sum_{k=1}^{n-1} T(X_k - A)T^\dagger \right\} \right\} \right\} \left\| \mathbb{E}_X\left\{ \exp\{T(X - A)T^\dagger\} \right\} \right\| \tag{A.50}$$

$$\leq \operatorname{Tr}\{I\} \left\| \mathbb{E}_X\left\{ \exp\{T(X - A)T^\dagger\} \right\} \right\|^n \tag{A.51}$$

$$= \dim(\mathcal{H}) \left\| \mathbb{E}_X\left\{ \exp\{T(X - A)T^\dagger\} \right\} \right\|^n \tag{A.52}$$

The first inequality follows from $\operatorname{Tr}\{AB\} \leq \operatorname{Tr}\{A\}\|B\|$. The second inequality follows from a repeated application of the same steps. The final equality follows because the trace of the identity operator is the dimension of the Hilbert space. This proves the "Bernstein trick" lemma. □

We finally come closer to proving the Operator Chernoff Bound. We first prove that the following inequality holds for IID operators X, X_1, $\ldots$, X_n such that $\mathbb{E}\{X\} \leq mI$, $A \geq aI$, and $1 \geq a \geq m \geq 0$:

$$\Pr\left\{ \sum_{k=1}^{n} X_k \not\leq nA \right\} \leq \dim(\mathcal{H}) \ \exp\{-nD(a||m)\}, \tag{A.53}$$

where $D(a||m)$ is the binary relative entropy:

$$D(a||m) = a\log a - a\log m + (1-a)\log(1-a) - (1-a)\log(1-m), \tag{A.54}$$

where the logarithm is the natural logarithm. We first apply the Bernstein Trick (Lemma A.1.3) with $T = \sqrt{t}I$:

$$\Pr\left\{ \sum_{k=1}^{n} X_k \not\leq nA \right\} \leq \Pr\left\{ \sum_{k=1}^{n} X_k \not\leq naI \right\} \tag{A.55}$$

$$\leq \dim(\mathcal{H}) \|\mathbb{E}\{\exp\{tX\}\exp\{-ta\}\}\|^n. \tag{A.56}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

So it is clear that it is best to optimize t in such a way that

$$\|\mathbb{E}\{\exp\{tX\}\exp\{-ta\}\}\| < 1 \tag{A.57}$$

so that we have exponential decay with increasing n. Now consider the following inequality:

$$\exp\{tX\} - I \leq X(\exp\{t\} - 1), \tag{A.58}$$

which holds because a similar one holds for all real $x \in (0, 1)$:

$$\frac{1}{x}(\exp\{tx\} - 1) \leq \exp\{t\} - 1. \tag{A.59}$$

Applying this inequality gives

$$\begin{aligned}
\mathbb{E}\{\exp\{tX\}\} &\leq \mathbb{E}\{X\}(\exp\{t\} - 1) + I && \text{(A.60)}\\
&\leq mI(\exp\{t\} - 1) + I && \text{(A.61)}\\
&= (m\exp\{t\} + 1 - m)I. && \text{(A.62)}
\end{aligned}$$

which in turn implies

$$\|\mathbb{E}\{\exp\{tX\}\exp\{-ta\}\}\| \leq (m\exp\{t\} + 1 - m)\exp\{-ta\}. \tag{A.63}$$

Choosing

$$t = \log\left(\frac{a}{m}\cdot\frac{1-m}{1-a}\right) > 0 \tag{A.64}$$

(which follows from the assumption that $a > m$) gives

$$\begin{aligned}
&(m\exp\{t\} + 1 - m)\exp\{-ta\}\\
&= \left(m\left(\frac{a}{m}\cdot\frac{1-m}{1-a}\right) + 1 - m\right)\exp\left\{-\log\left(\frac{a}{m}\cdot\frac{1-m}{1-a}\right)a\right\} && \text{(A.65)}\\
&= \left(a\cdot\frac{1-m}{1-a} + 1 - m\right)\exp\left\{-a\log\left(\frac{a}{m}\right) - a\log\left(\frac{1-m}{1-a}\right)\right\} && \text{(A.66)}\\
&= \left(\frac{1-m}{1-a}\right)\exp\left\{-a\log\left(\frac{a}{m}\right) - a\log\left(\frac{1-m}{1-a}\right)\right\} && \text{(A.67)}\\
&= \exp\left\{-a\log\left(\frac{a}{m}\right) - (1-a)\log\left(\frac{1-a}{1-m}\right)\right\} && \text{(A.68)}\\
&= \exp\{-D(a \mid\mid m)\}, && \text{(A.69)}
\end{aligned}$$

proving the desired bound in (A.53).

By substituting $Y_k = I - X_k$ and $B = I - A$ into (A.53) and having the opposite conditions $\mathbb{E}\{X\} \geq mI$, $A \leq aI$, and $0 \leq a \leq m \leq 1$, we can show that the following inequality holds for IID operators X, X_1, $\ldots$, X_n:

$$\Pr\left\{\sum_{k=1}^{n} X_k \not\geq nA\right\} \leq \dim(\mathcal{H})\ \exp\{-nD(a||m)\}. \tag{A.70}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

To finish off the proof of the Operator Chernoff Bound, consider the variables $Z_i = M\mu^{-1/2}X_i\mu^{-1/2}$ with $\mu \equiv \mathbb{E}\{X\} \geq MI$. Then $\mathbb{E}\{Z_i\} = MI$ and $0 \leq Z_i \leq I$. The following events are thus equivalent

$$(1-\eta)\mu \leq \frac{1}{n}\sum_{i=1}^{n} X_i \leq (1+\eta)\mu \iff (1-\eta)MI \leq \frac{1}{n}\sum_{i=1}^{n} Z_i \leq (1+\eta)MI, \tag{A.71}$$

and we can apply (A.53), (A.70), and the union bound to obtain

$$\Pr\left\{\left((1-\eta)\mu \not\leq \frac{1}{n}\sum_{i=1}^{n} X_i\right) \bigcup \left(\frac{1}{n}\sum_{i=1}^{n} X_i \not\leq (1+\eta)\mu\right)\right\} \tag{A.72}$$

$$\leq \dim(\mathcal{H})\ \exp\{-nD((1-\eta)M||M)\} + \dim(\mathcal{H})\ \exp\{-nD((1+\eta)M||M)\} \tag{A.73}$$

$$\leq 2\dim(\mathcal{H})\ \exp\left\{-n\ \frac{\eta^2 M}{4\ln 2}\right\}, \tag{A.74}$$

where the last line exploits the following inequality valid for $-1/2 \leq \eta \leq 1/2$ and $(1+\eta)M \leq 1$:

$$D((1+\eta)M||M) \geq \frac{1}{4\ln 2}\eta^2 M. \tag{A.75}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

APPENDIX B

Monotonicity of Quantum Relative Entropy

The following proof of monotonicity of quantum relative entropy is due to Nielsen and Petz [198].

Theorem B.0.1. *For any two bipartite quantum states ρ^{XY} and σ^{XY}, the quantum relative entropy is monotone under the discarding of systems:*

$$D\big(\rho^{XY} \mid\mid \sigma^{XY}\big) \geq D\big(\rho^{X} \mid\mid \sigma^{X}\big). \tag{B.1}$$

Proof. We first require the notion of operator convexity. Recall the partial order $A \geq B$ for Hermitian operators A and B where $A \geq B$ if $A - B$ is a positive operator. A function f is operator convex if for all n, for all $A, B \in M_n$ (where M_n is the set of $n \times n$ Hermitian operators), and for all $\lambda \in [0, 1]$, we have

$$f(\lambda A + (1-\lambda)B) \leq \lambda f(A) + (1-\lambda) f(B). \tag{B.2}$$

We require Lemma B.0.4, that states that $-\ln(x)$ is an operator convex function and Lemma B.0.5, that states that if f is an operator convex function and $U^{V\to W}$ is an isometry (where $\dim(V) \leq \dim(W)$), then $f\big(U^\dagger X U\big) \leq U^\dagger f(X) U$ for all operators X. We need to reexpress the quantum relative entropy using a linear map on matrices known as the *relative modular operator*. We are assuming that ρ and σ are invertible when defining this operator, but the proof technique here extends with a continuity argument. Let $\mathcal{L}(A) \equiv \sigma A$ and $\mathcal{R}(A) \equiv A\rho^{-1}$. The relative modular operator Δ is the product of these linear maps under composition: $\Delta(A) \equiv \mathcal{L}(\mathcal{R}(A))$. The two superoperators $\mathcal{L}$ and $\mathcal{R}$ commute so that $\Delta(A) \equiv \mathcal{R}(\mathcal{L}(A))$. We now define a function ln on superoperators $\mathcal{E}$, where $\mathcal{E}$ is a linear map that is strictly positive with respect to the Hilbert-Schmidt inner product $\langle A, B\rangle \equiv \mathrm{Tr}\big\{A^\dagger B\big\}$. To define this function, we expand the linear map $\mathcal{E}$ in a basis where it is diagonal: $\mathcal{E} = \sum_x \mu_x \mathcal{E}_x$, so that $\ln(\mathcal{E}) \equiv \sum_x \ln(\mu_x)\mathcal{E}_x$. Observe now that $\mathcal{L}(A)$, $\mathcal{R}(A)$, and

$\Delta(A)$ are all strictly positive for $A \neq 0$ because

$$\langle A, \mathcal{L}(A)\rangle = \mathrm{Tr}\{A^\dagger \sigma A\} = \mathrm{Tr}\{AA^\dagger \sigma\} > 0, \tag{B.3}$$
$$\langle A, \mathcal{R}(A)\rangle = \mathrm{Tr}\{A^\dagger A \rho^{-1}\} > 0, \tag{B.4}$$
$$\langle A, \Delta(A)\rangle = \mathrm{Tr}\{A^\dagger \sigma A \rho^{-1}\} > 0, \tag{B.5}$$

and our assumption that ρ and σ are positive, invertible operators. After expanding the maps $\mathcal{L}(A)$ and $\mathcal{R}(A)$ in a diagonal basis, observe that $\ln(\mathcal{L})(A) = \ln(\sigma)A$ and $\ln(\mathcal{R})(A) = -A\ln(\rho)$. Also, we have that

$$\ln(\Delta) = \ln(\mathcal{L}) + \ln(\mathcal{R}) \tag{B.6}$$

because $\mathcal{L}$ and $\mathcal{R}$ commute. We can then rewrite the quantum relative entropy as follows:

$$D(\rho \,||\, \sigma) = \mathrm{Tr}\{\rho \ln \rho - \rho \ln \sigma\} \tag{B.7}$$
$$= \mathrm{Tr}\{\rho \ln \rho\} - \mathrm{Tr}\{\rho \ln \sigma\} \tag{B.8}$$
$$= \mathrm{Tr}\{\rho^{1/2}\rho^{1/2} \ln \rho\} - \mathrm{Tr}\{\rho^{1/2}\rho^{1/2} \ln \sigma\} \tag{B.9}$$
$$= -\mathrm{Tr}\{\rho^{1/2} \ln(\mathcal{R})(\rho^{1/2})\} - \mathrm{Tr}\{\ln(\mathcal{L})(\rho^{1/2})\rho^{1/2}\} \tag{B.10}$$
$$= \mathrm{Tr}\{\rho^{1/2}[-\ln(\mathcal{R})(\rho^{1/2}) - \ln(\mathcal{L})(\rho^{1/2})]\} \tag{B.11}$$
$$= \langle \rho^{1/2}, -\ln(\Delta)(\rho^{1/2})\rangle. \tag{B.12}$$

The statement of monotonicity of quantum relative entropy then becomes

$$\left\langle (\rho^X)^{1/2}, -\ln(\Delta^X)\left((\rho^X)^{1/2}\right)\right\rangle \leq \left\langle (\rho^{XY})^{1/2}, -\ln(\Delta^{XY})\left((\rho^{XY})^{1/2}\right)\right\rangle, \tag{B.13}$$

where

$$\Delta^X(A) \equiv \sigma^X A(\rho^X)^{-1}, \tag{B.14}$$
$$\Delta^{XY}(A) \equiv \sigma^{XY} A(\rho^{XY})^{-1}. \tag{B.15}$$

To complete the proof, suppose that an isometry U from X to XY has the following properties:

$$U^\dagger \Delta^{XY} U = \Delta^X, \tag{B.16}$$
$$U\left((\rho^X)^{1/2}\right) = (\rho^{XY})^{1/2}. \tag{B.17}$$

We investigate the consequences of the existence of such an isometry and later explicitly construct it. First, we rewrite monotonicity of quantum relative entropy in terms of this isometry:

$$\left\langle (\rho^X)^{1/2}, -\ln(U^\dagger \Delta^{XY} U)\left((\rho^X)^{1/2}\right)\right\rangle \leq \left\langle (\rho^{XY})^{1/2}, -\ln(\Delta^{XY})\left((\rho^{XY})^{1/2}\right)\right\rangle. \tag{B.18}$$

Now we know from Lemmas B.0.4 and B.0.5 on operator convexity that

$$-\ln(U^\dagger \Delta^{XY} U) \leq -U^\dagger \ln(\Delta^{XY}) U, \tag{B.19}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

so that

$$\begin{aligned}
&\left\langle \left(\rho^X\right)^{1/2}, -\ln\left(U^\dagger \Delta^{XY} U\right)\left(\left(\rho^X\right)^{1/2}\right)\right\rangle \\
&\leq \left\langle \left(\rho^X\right)^{1/2}, -U^\dagger \ln\left(\Delta^{XY}\right) U\left(\left(\rho^X\right)^{1/2}\right)\right\rangle && \text{(B.20)} \\
&= \left\langle U\left(\left(\rho^X\right)^{1/2}\right), -\ln\left(\Delta^{XY}\right) U\left(\left(\rho^X\right)^{1/2}\right)\right\rangle && \text{(B.21)} \\
&= \left\langle \left(\rho^{XY}\right)^{1/2}, -\ln\left(\Delta^{XY}\right)\left(\left(\rho^{XY}\right)^{1/2}\right)\right\rangle. && \text{(B.22)}
\end{aligned}$$

This completes the proof of monotonicity of quantum relative entropy if it is true that there exists an isometry satisfying the properties in (B.16-B.17). Consider the following choice for the map U:

$$U(A) \equiv \left(A\left(\rho^X\right)^{-1/2} \otimes I^Y\right)\left(\rho^{XY}\right)^{1/2}. \tag{B.23}$$

This choice for U satisfies (B.17) by inspection. The adjoint $U^\dagger$ is some operator satisfying

$$\langle B, U(A)\rangle = \left\langle U^\dagger(B), A\right\rangle \tag{B.24}$$

for all $A \in \mathcal{H}^X$ and $B \in \mathcal{H}^{XY}$. Thus, we require some operator $U^\dagger$ such that

$$\begin{aligned}
\operatorname{Tr}\left\{\left[U^\dagger(B)\right]^\dagger A\right\} &= \left\langle U^\dagger(B), A\right\rangle && \text{(B.25)} \\
&= \langle B, U(A)\rangle && \text{(B.26)} \\
&= \operatorname{Tr}\left\{B^\dagger\left(A\left(\rho^X\right)^{-1/2} \otimes I^Y\right)\left(\rho^{XY}\right)^{1/2}\right\} && \text{(B.27)} \\
&= \operatorname{Tr}\left\{\left(\left(\rho^X\right)^{-1/2} \otimes I^Y\right)\left(\rho^{XY}\right)^{1/2} B^\dagger A\right\}. && \text{(B.28)}
\end{aligned}$$

So the adjoint $U^\dagger$ is as follows:

$$U^\dagger(B) = \operatorname{Tr}_Y\left\{B\left(\rho^{XY}\right)^{1/2}\left(\left(\rho^X\right)^{-1/2} \otimes I^Y\right)\right\}. \tag{B.29}$$

We can now verify that (B.16) holds

$$\begin{aligned}
&U^\dagger \Delta^{XY} U(A) \\
&= \operatorname{Tr}_Y\left\{\left\{\sigma^{XY}\left[\left(A\left(\rho^X\right)^{-1/2} \otimes I^Y\right)\left(\rho^{XY}\right)^{1/2}\right]\left(\rho^{XY}\right)^{-1}\right\}\left(\rho^{XY}\right)^{1/2}\left(\left(\rho^X\right)^{-1/2} \otimes I^Y\right)\right\} && \text{(B.30)} \\
&= \operatorname{Tr}_Y\left\{\left\{\sigma^{XY}\left[\left(A\left(\rho^X\right)^{-1/2} \otimes I^Y\right) I^{XY}\right]\right\}\left(\left(\rho^X\right)^{-1/2} \otimes I^Y\right)\right\} && \text{(B.31)} \\
&= \operatorname{Tr}_Y\left\{\sigma^{XY} A\left(\rho^X\right)^{-1} \otimes I^Y\right\} && \text{(B.32)} \\
&= \sigma^X A\left(\rho^X\right)^{-1} && \text{(B.33)} \\
&= \Delta^X(A). && \text{(B.34)}
\end{aligned}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The map U is also an isometry because

$$U^\dagger(U(A)) = \text{Tr}_Y\left\{\left(A\left(\rho^X\right)^{-1/2}\otimes I^Y\right)\left(\rho^{XY}\right)^{1/2}\left(\rho^{XY}\right)^{1/2}\left(\left(\rho^X\right)^{-1/2}\otimes I^Y\right)\right\} \tag{B.35}$$

$$= \text{Tr}_Y\left\{\left(A\left(\rho^X\right)^{-1/2}\otimes I^Y\right)\rho^{XY}\left(\left(\rho^X\right)^{-1/2}\otimes I^Y\right)\right\} \tag{B.36}$$

$$= \left(A\left(\rho^X\right)^{-1/2}\otimes I^Y\right)\rho^X\left(\left(\rho^X\right)^{-1/2}\otimes I^Y\right) \tag{B.37}$$

$$= A. \tag{B.38}$$

This concludes the proof of the monotonicity of quantum relative entropy. Observe that the main tool exploited in proving this theorem is the operator convexity of the function $-\ln(x)$. □

Lemma B.0.4. $-\ln(x)$ *is an operator convex function.*

Proof. We begin by proving that x^{-1} is operator convex on $(0,\infty)$. Let A and B be strictly positive operators such that $A \leq B$. Begin with the special case where $A = I$ and the goal then becomes to prove that

$$(\lambda I + (1-\lambda)B)^{-1} \leq \lambda I + (1-\lambda)B^{-1}. \tag{B.39}$$

The above result follows because I and B commute and the function x^{-1} is convex on the real numbers. Now make the substitution $B \to A^{-1/2}BA^{-1/2}$ in the above to obtain

$$\left(\lambda I + (1-\lambda)A^{-1/2}BA^{-1/2}\right)^{-1} \leq \lambda I + (1-\lambda)\left(A^{-1/2}BA^{-1/2}\right)^{-1}. \tag{B.40}$$

Conjugating by $A^{-1/2}$ gives the desired inequality:

$$A^{-1/2}\left(\lambda I + (1-\lambda)A^{-1/2}BA^{-1/2}\right)^{-1}A^{-1/2}$$
$$\leq A^{-1/2}\left(\lambda I + (1-\lambda)\left(A^{-1/2}BA^{-1/2}\right)^{-1}\right)A^{-1/2}, \tag{B.41}$$

$$\therefore (\lambda A + (1-\lambda)B)^{-1} \leq \lambda A^{-1} + (1-\lambda)B^{-1}. \tag{B.42}$$

We can now prove the operator convexity of $-\ln(A)$ by exploiting the above result and the following integral representation of $-\ln(a)$:

$$-\ln(a) = \int_0^\infty dt\ \frac{1}{a+t} - \frac{1}{1-t}. \tag{B.43}$$

The following integral representation then follows for a strictly positive operator A:

$$-\ln(A) = \int_0^\infty dt\ (A+tI)^{-1} - (I+tI)^{-1}. \tag{B.44}$$

Operator convexity of $-\ln(A)$ follows if $(A+tI)^{-1}$ is operator convex:

$$(\lambda A + (1-\lambda)B + tI)^{-1} \leq \lambda(A+tI)^{-1} + (1-\lambda)(B+tI)^{-1}. \tag{B.45}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

The above statement follows by rewriting the LHS as

$$(\lambda(A+tI)+(1-\lambda)(B+tI))^{-1} \tag{B.46}$$

and applying operator convexity of x^{-1}. □

Lemma B.0.5. *If f is an operator convex function and $U^{V\to W}$ is an isometry (where $\dim(V)\leq\dim(W)$), then $f\left(U^\dagger XU\right)\leq U^\dagger f(X)U$ for all operators X.*

Proof. First recall that $f\left(U^\dagger XU\right)=U^\dagger f(X)U$ if U maps V onto W. This follows from the way that we apply functions to operators. The more difficult part of the proof is proving the inequality if U does not map V onto W. Let Π be a projector onto the range W' of U (W' is the subspace of W into which the isometry takes vectors in V), and let $\hat{\Pi}\equiv I-\Pi$ be a projector onto the orthocomplement. It is useful to adopt the notation f_V, $f_{W'}$, and f_W to denote the three different spaces on which the function f can act. Observe that $\Pi U=U$ because Π projects onto the range of U. It follows that

$$f_V\left(U^\dagger XU\right)=f_V\left(U^\dagger\Pi(\Pi X\Pi)\Pi U\right), \tag{B.47}$$

and we can then conclude that

$$f_V\left(U^\dagger\Pi(\Pi X\Pi)\Pi U\right)=U^\dagger\Pi f_{W'}(\Pi X\Pi)\Pi U, \tag{B.48}$$

from our observation at the beginning of the proof. It then suffices to prove the following inequality in order to prove the inequality in the statement of the lemma:

$$f_{W'}(\Pi X\Pi)\leq\Pi f_W(X)\Pi. \tag{B.49}$$

Consider that

$$f_{W'}(\Pi X\Pi)=\Pi f_W(\Pi X\Pi)\Pi=\Pi f_W\left(\Pi X\Pi+\hat{\Pi}X\hat{\Pi}\right)\Pi, \tag{B.50}$$

because

$$f_W\left(\Pi X\Pi+\hat{\Pi}X\hat{\Pi}\right)=f_W(\Pi X\Pi)+f_W\left(\hat{\Pi}X\hat{\Pi}\right), \tag{B.51}$$

$$\Pi f_W\left(\hat{\Pi}X\hat{\Pi}\right)\Pi=0. \tag{B.52}$$

Let $S\equiv\Pi-\hat{\Pi}$ be a unitary on W and recalling that $\Pi+\hat{\Pi}=I$, it follows that

$$\frac{X+SXS^\dagger}{2}=\frac{\left(\Pi+\hat{\Pi}\right)X\left(\Pi+\hat{\Pi}\right)+\left(\Pi-\hat{\Pi}\right)X\left(\Pi-\hat{\Pi}\right)}{2} \tag{B.53}$$

$$=\Pi X\Pi+\hat{\Pi}X\hat{\Pi}. \tag{B.54}$$

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

We can then apply the above equality and operator convexity of f to give

$$f_W\left(\Pi X \Pi + \hat{\Pi} X \hat{\Pi}\right) = f_W\left(\frac{X + SXS^\dagger}{2}\right) \tag{B.55}$$

$$\leq \frac{1}{2}\left[f_W(X) + f_W\left(SXS^\dagger\right)\right] \tag{B.56}$$

$$= \frac{1}{2}\left[f_W(X) + S f_W(X) S^\dagger\right] \tag{B.57}$$

$$= \Pi f_X(X) \Pi + \hat{\Pi} f_X(X) \hat{\Pi}. \tag{B.58}$$

Conjugating by Π and recalling (B.50) gives the desired inequality that is sufficient to prove the one in the statement of the lemma:

$$\Pi f_W\left(\Pi X \Pi + \hat{\Pi} X \hat{\Pi}\right) \Pi \leq \Pi f_X(X) \Pi. \tag{B.59}$$

□

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Bibliography

[1] William Thomson (1st Baron Kelvin). Nineteenth-century clouds over the dynamical theory of heat and light. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 2(6):1, 1901.

[2] Anura Abeyesinghe. *Unification of Quantum Information Theory*. PhD thesis, California Institute of Technology, 2006.

[3] Anura Abeyesinghe, Igor Devetak, Patrick Hayden, and Andreas Winter. The mother of all protocols: Restructuring quantum information's family tree. *Proceedings of the Royal Society A*, 465(2108):2537–2563, August 2009. arXiv:quant-ph/0606225.

[4] Anura Abeyesinghe and Patrick Hayden. Generalized remote state preparation: Trading cbits, qubits, and ebits in quantum communication. *Physical Review A*, 68(6):062319, December 2003.

[5] Christoph Adami and Nicolas J. Cerf. von Neumann capacity of noisy quantum channels. *Physical Review A*, 56(5):3470–3483, November 1997.

[6] Dorit Aharonov and Michael Ben-Or. Fault-tolerant quantum computation with constant error. In *STOC '97: Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, pages 176–188, New York, NY, USA, 1997. ACM.

[7] Rudolph Ahlswede and Andreas J. Winter. Strong converse for identification via quantum channels. *IEEE Transactions in Information Theory*, 48(3):569–579, March 2002. arXiv:quant-ph/0012127.

[8] Charlene Ahn, Andrew Doherty, Patrick Hayden, and Andreas Winter. On the distributed compression of quantum information. *IEEE Transactions on Information Theory*, 52(10):4349–4357, October 2006.

[9] Robert Alicki and Mark Fannes. Continuity of quantum conditional information. *Journal of Physics A: Mathematical and General*, 37(5):L55–L57, 2004.

[10] Alain Aspect, Philippe Grangier, and Gérard Roger. Experimental tests of realistic local theories via Bell's theorem. *Physical Review Letters*, 47(7):460–463, August 1981.

[11] Guillaume Aubrun, Stanislaw Szarek, and Elisabeth Werner. Hastings' additivity counterexample via Dvoretzky's theorem. *Communications in Mathematical Physics*, 305(1):85–97, 2011. arXiv:1003.4925.

[12] Koenraad M. R. Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127, 2007.

[13] Howard Barnum, Carlton M. Caves, Christopher A. Fuchs, Richard Jozsa, and Benjamin Schumacher. On quantum coding for ensembles of mixed states. *Journal of Physics A: Mathematical and General*, 34(35):6767, 2001.

[14] Howard Barnum, Patrick Hayden, Richard Jozsa, and Andreas Winter. On the reversible extraction of classical information from a quantum source. *Proceedings of the Royal Society A*, 457(2012):2019–2039, July 2001.

[15] Howard Barnum, Emanuel Knill, and Michael A. Nielsen. On quantum fidelities and channel capacities. *IEEE Transactions on Information Theory*, 46:1317–1329, 2000.

[16] Howard Barnum, M. A. Nielsen, and Benjamin Schumacher. Information transmission through a noisy quantum channel. *Physical Review A*, 57(6):4153–4175, June 1998.

[17] John Stewart Bell. On the Einstein-Podolsky-Rosen paradox. *Physics*, 1:195–200, 1964.

[18] Charles H. Bennett. Quantum cryptography using any two nonorthogonal states. *Physical Review Letters*, 68(21):3121–3124, May 1992.

[19] Charles H. Bennett. Quantum information and computation. *Physics Today*, 48(10):24–30, October 1995.

[20] Charles H. Bennett. A resource-based view of quantum information. *Quantum Information and Computation*, 4:460–466, December 2004.

[21] Charles H. Bennett, Herbert J. Bernstein, Sandu Popescu, and Benjamin Schumacher. Concentrating partial entanglement by local operations. *Physical Review A*, 53(4):2046–2052, April 1996.

[22] Charles H. Bennett and Gilles Brassard. Quantum cryptography: Public key distribution and coin tossing. In *Proceedings of IEEE International Conference on Computers Systems and Signal Processing*, pages 175–179, Bangalore, India, December 1984.

[23] Charles H. Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K. Wootters. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Physical Review Letters*, 70(13):1895–1899, Mar 1993.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[24] Charles H. Bennett, Gilles Brassard, and Artur K. Ekert. Quantum cryptography. *Scientific American*, pages 50–57, October 1992.

[25] Charles H. Bennett, Gilles Brassard, and N. David Mermin. Quantum cryptography without Bell's theorem. *Physical Review Letters*, 68(5):557–559, February 1992.

[26] Charles H. Bennett, Gilles Brassard, Sandu Popescu, Benjamin Schumacher, John A. Smolin, and William K. Wootters. Purification of noisy entanglement and faithful teleportation via noisy channels. *Physical Review Letters*, 76(5):722–725, January 1996.

[27] Charles H. Bennett, Igor Devetak, Aram W. Harrow, Peter W. Shor, and Andreas Winter. Quantum reverse Shannon theorem. December 2009. arXiv:0912.5537.

[28] Charles H. Bennett, David P. DiVincenzo, Peter W. Shor, John A. Smolin, Barbara M. Terhal, and William K. Wootters. Remote state preparation. *Physical Review Letters*, 87(7):077902, July 2001.

[29] Charles H. Bennett, David P. DiVincenzo, and John A. Smolin. Capacities of quantum erasure channels. *Physical Review Letters*, 78(16):3217–3220, April 1997.

[30] Charles H. Bennett, David P. DiVincenzo, John A. Smolin, and William K. Wootters. Mixed-state entanglement and quantum error correction. *Physical Review A*, 54(5):3824–3851, Nov 1996.

[31] Charles H. Bennett, Aram W. Harrow, and Seth Lloyd. Universal quantum data compression via nondestructive tomography. *Physical Review A*, 73(3):032336, March 2006.

[32] Charles H. Bennett, Patrick Hayden, Debbie W. Leung, Peter W. Shor, and Andreas Winter. Remote preparation of quantum states. *IEEE Transactions on Information Theory*, 51(1):56–74, January 2005.

[33] Charles H. Bennett, Peter W. Shor, John A. Smolin, and Ashish V. Thapliyal. Entanglement-assisted classical capacity of noisy quantum channels. *Physical Review Letters*, 83(15):3081–3084, October 1999.

[34] Charles H. Bennett, Peter W. Shor, John A. Smolin, and Ashish V. Thapliyal. Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem. *IEEE Transactions on Information Theory*, 48:2637–2655, 2002.

[35] Charles H. Bennett and Stephen J. Wiesner. Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states. *Physical Review Letters*, 69(20):2881–2884, Nov 1992.

[36] Toby Berger. *Rate distortion theory: A mathematical basis for data compression*. Prentice-Hall, Englewood Cliffs, New Jersey, USA, 1971.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[37] Toby Berger. Multiterminal source coding. *The Information Theory Approach to Communications*, 1977. Springer-Verlag, New York.

[38] Mario Berta, Matthias Christandl, Roger Colbeck, Joseph M. Renes, and Renato Renner. The uncertainty principle in the presence of quantum memory. *Nature Physics*, 6:659–662, 2010. arXiv:0909.0950.

[39] Mario Berta, Matthias Christandl, and Renato Renner. The quantum reverse Shannon theorem based on one-shot information theory. *Communications in Mathematical Physics*, 306(3):579–615, August 2011. arXiv:0912.3805.

[40] Robin Blume-Kohout, Sarah Croke, and Daniel Gottesman. Streaming universal distortion-free entanglement concentration. October 2009. arXiv:0910.5952.

[41] David Bohm. *Quantum Theory*. Courier Dover Publications, 1989.

[42] Garry Bowen. Quantum feedback channels. *IEEE Transactions in Information Theory*, 50(10):2429–2434, October 2004. arXiv:quant-ph/0209076.

[43] Garry Bowen. Feedback in quantum communication. *International Journal of Quantum Information*, 3(1):123–127, 2005. arXiv:quant-ph/0410191.

[44] Garry Bowen and Rajagopal Nagarajan. On feedback and the classical capacity of a noisy quantum channel. *IEEE Transactions in Information Theory*, 51(1):320–324, January 2005. arXiv:quant-ph/0305176.

[45] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, The Edinburgh Building, Cambridge, CB2 8RU, UK, 2004.

[46] Kamil Brádler, Patrick Hayden, Dave Touchette, and Mark M. Wilde. Trade-off capacities of the quantum Hadamard channels. *Physical Review A*, 81(6):062312, June 2010.

[47] Fernando G.S.L. Brandao and Michal Horodecki. On Hastings' counterexamples to the minimum output entropy additivity conjecture. *Open Systems & Information Dynamics*, 17(1):31–52, 2010. arXiv:0907.3210.

[48] Samuel L. Braunstein, Christopher A. Fuchs, Daniel Gottesman, and Hoi-Kwong Lo. A quantum analog of Huffman coding. *IEEE Transactions in Information Theory*, 46(4):1644–1649, July 2000.

[49] Todd A. Brun. Quantum information processing course lecture slides. http://almaak.usc.edu/ tbrun/Course/.

[50] Francesco Buscemi and Nilanjana Datta. The quantum capacity of channels with arbitrarily correlated noise. *IEEE Transactions in Information Theory*, 56(3):1447–1460, March 2010. arXiv:0902.0158.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[51] N. Cai, Andreas Winter, and Raymond W. Yeung. Quantum privacy and quantum wiretap channels. *Problems of Information Transmission*, 40(4):318–336, October 2004.

[52] A. Robert Calderbank, Eric M. Rains, Peter W. Shor, and N. J. A. Sloane. Quantum error correction and orthogonal geometry. *Physical Review Letters*, 78(3):405–408, January 1997.

[53] A. Robert Calderbank, Eric M. Rains, Peter W. Shor, and N. J. A. Sloane. Quantum error correction via codes over GF(4). *IEEE Transactions on Information Theory*, 44:1369–1387, 1998.

[54] A. Robert Calderbank and Peter W. Shor. Good quantum error-correcting codes exist. *Physical Review A*, 54(2):1098–1105, August 1996.

[55] N. J. Cerf and C. Adami. Negative entropy and information in quantum mechanics. *Physical Review Letters*, 79:5194–5197, 1997.

[56] Patrick J. Coles, Roger Colbeck, Li Yu, and Michael Zwolak. Uncertainty relations from simple entropic properties. December 2011. arXiv:1112.0543.

[57] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[58] Imre Csiszár and Janos Korner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Probability and mathematical statistics. Akademiai Kiado, Budapest, 1981. Out of print.

[59] L. Czekaj and P. Horodecki. Nonadditivity effects in classical capacities of quantum multiple-access channels. *arXiv:0807.3977*, July 2008.

[60] Nilanjana Datta. Min- and max-relative entropies and a new entanglement monotone. *IEEE Transactions on Information Theory*, 55(6):2816–2826, June 2009. arXiv:0803.2770.

[61] Nilanjana Datta and Min-Hsiu Hsieh. Universal coding for transmission of private information. *Journal of Mathematical Physics*, 51(12):122202, 2010. arXiv:1007.2629.

[62] Nilanjana Datta and Min-Hsiu Hsieh. The apex of the family tree of protocols: Optimal rates and resource inequalities. *New Journal of Physics*, 13:093042, September 2011. arXiv:1103.1135.

[63] Nilanjana Datta and Min-Hsiu Hsieh. One-shot entanglement-assisted classical communication. May 2011. arXiv:1105.3321.

[64] Nilanjana Datta and Renato Renner. Smooth entropies and the quantum information spectrum. *IEEE Transactions on Information Theory*, 55(6):2807–2815, June 2009. arXiv:0801.0282.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[65] E. B. Davies and J. T. Lewis. An operational approach to quantum probability. *Communications in Mathematical Physics*, 17(3):239–260, 1970.

[66] Louis de Broglie. *Recherches sur la théorie des quanta*. PhD thesis, Paris, 1924.

[67] David Deutsch. Quantum theory, the Church-Turing principle and the universal quantum computer. *Proceedings of the Royal Society of London A*, 400(1818):97–117, July 1985.

[68] Igor Devetak. The private classical capacity and quantum capacity of a quantum channel. *IEEE Transactions on Information Theory*, 51:44–55, January 2005.

[69] Igor Devetak. Triangle of dualities between quantum communication protocols. *Physical Review Letters*, 97(14):140503, October 2006.

[70] Igor Devetak, Aram W. Harrow, and Andreas Winter. A resource framework for quantum Shannon theory. *IEEE Transactions on Information Theory*, 54(10):4587–4618, October 2008.

[71] Igor Devetak, Aram W. Harrow, and Andreas J. Winter. A family of quantum protocols. *Physical Review Letters*, 93:239503, 2004.

[72] Igor Devetak, Marius Junge, Christopher King, and Mary Beth Ruskai. Multiplicativity of completely bounded p-norms implies a new additivity result. *Communications in Mathematical Physics*, 266(1):37–63, 2006.

[73] Igor Devetak and Peter W. Shor. The capacity of a quantum channel for simultaneous transmission of classical and quantum information. *Communications in Mathematical Physics*, 256:287–303, 2005.

[74] Igor Devetak and Andreas Winter. Classical data compression with quantum side information. *Physical Review A*, 68(4):042301, October 2003.

[75] Igor Devetak and Andreas Winter. Relating quantum privacy and quantum coherence: An operational approach. *Physical Review Letters*, 93(8):080501, August 2004.

[76] Igor Devetak and Andreas Winter. Distillation of secret key and entanglement from quantum states. *Proceedings of the Royal Society A*, 461:207–235, 2005.

[77] Igor Devetak and Jon Yard. Exact cost of redistributing multipartite quantum states. *Physical Review Letters*, 100(23):230501, June 2008.

[78] D. Dieks. Communication by EPR devices. *Physics Letters A*, 92:271, 1982.

[79] P. A. M. Dirac. *The Principles of Quantum Mechanics (International Series of Monographs on Physics)*. Oxford University Press, USA, February 1982.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[80] David P. DiVincenzo, Peter W. Shor, and John A. Smolin. Quantum-channel capacity of very noisy channels. *Physical Review A*, 57(2):830–839, February 1998.

[81] Jonathan P. Dowling and Gerard J Milburn. Quantum technology: The second quantum revolution. *Philosophical Transactions of The Royal Society of London Series A*, 361(1809):1655–1674, August 2003.

[82] Frederic Dupuis. *The decoupling approach to quantum information theory*. PhD thesis, University of Montreal, April 2010. arXiv:1004.1641.

[83] Frederic Dupuis, Mario Berta, Jürg Wullschleger, and Renato Renner. The decoupling theorem. December 2010. arXiv:1012.6044.

[84] Frederic Dupuis, Patrick Hayden, and Ke Li. A father protocol for quantum broadcast channels. *IEEE Transations on Information Theory*, 56(6):2946–2956, June 2010. arXiv:quant-ph/0612155.

[85] Nicolas Dutil. *Multiparty quantum protocols for assisted entanglement distillation*. PhD thesis, McGill University, May 2011. arXiv:1105.4657.

[86] Albert Einstein. Über einen die erzeugung und verwandlung des lichtes betreffenden heuristischen gesichtspunkt. *Annalen der Physik*, 17:132–148, 1905.

[87] Albert Einstein, Boris Podolsky, and Nathan Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47:777–780, 1935.

[88] Artur K. Ekert. Quantum cryptography based on Bell's theorem. *Physical Review Letters*, 67(6):661–663, August 1991.

[89] Abbas El Gamal and Young-Han Kim. Lecture notes on network information theory. January 2010. arXiv:1001.3404.

[90] Peter Elias. The efficient construction of an unbiased random sequence. *Annals of Mathematical Statistics*, 43(3):865–870, 1972.

[91] Mark Fannes. A continuity property of the entropy density for spin lattices. *Communications in Mathematical Physics*, 31:291, 1973.

[92] Robert Mario Fano. Fano inequality. *Scholarpedia*, 3(10):6648, 2008.

[93] Omar Fawzi, Patrick Hayden, Ivan Savov, Pranab Sen, and Mark M. Wilde. Classical communication over a quantum interference channel. *Accepted into IEEE Transactions on Information Theory*, February 2011. arXiv:1102.2624.

[94] Richard P. Feynman. Simulating physics with computers. *International Journal of Theoretical Physics*, 21:467–488, 1982.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[95] Richard P. Feynman. *Feynman Lectures On Physics (3 Volume Set)*. Addison Wesley Longman, September 1998.

[96] Christopher Fuchs. *Distinguishability and Accessible Information in Quantum Theory*. PhD thesis, University of New Mexico, December 1996. arXiv:quant-ph/9601020.

[97] Christopher A. Fuchs and Jeroen van de Graaf. Cryptographic distinguishability measures for quantum mechanical states. *IEEE Transactions on Information Theory*, 45(4):1216–1227, May 1998. arXiv:quant-ph/9712042.

[98] Motohisa Fukuda and Christopher King. Entanglement of random subspaces via the Hastings bound. *Journal of Mathematical Physics*, 51(4):042201, 2010.

[99] Motohisa Fukuda, Christopher King, and David K. Moser. Comments on Hastings' additivity counterexamples. *Communications in Mathematical Physics*, 296(1):111–143, 2010. arXiv:0905.3697.

[100] Raúl García-Patrón, Stefano Pirandola, Seth Lloyd, and Jeffrey H. Shapiro. Reverse coherent information. *Physical Review Letters*, 102(21):210501, May 2009.

[101] Walther Gerlach and Otto Stern. Das magnetische moment des silberatoms. *Zeitschrift für Physik*, 9:353–355, 1922.

[102] Vittorio Giovannetti and Rosario Fazio. Information-capacity description of spin-chain correlations. *Physical Review A*, 71(3):032314, March 2005.

[103] Vittorio Giovannetti, Saikat Guha, Seth Lloyd, Lorenzo Maccone, and Jeffrey H. Shapiro. Minimum output entropy of bosonic channels: A conjecture. *Physical Review A*, 70(3):032315, September 2004.

[104] Vittorio Giovannetti, Saikat Guha, Seth Lloyd, Lorenzo Maccone, Jeffrey H. Shapiro, and Horace P. Yuen. Classical capacity of the lossy bosonic channel: The exact solution. *Physical Review Letters*, 92(2):027902, January 2004.

[105] Vittorio Giovannetti, Alexander S. Holevo, Seth Lloyd, and Lorenzo Maccone. Generalized minimal output entropy conjecture for one-mode Gaussian channels: definitions and some exact results. *Journal of Physics A: Mathematical and Theoretical*, 43(41):415305, 2010.

[106] Vittorio Giovannetti, Seth Lloyd, Lorenzo Maccone, and Peter W. Shor. Broadband channel capacities. *Physical Review A*, 68(6):062323, December 2003.

[107] Vittorio Giovannetti, Seth Lloyd, Lorenzo Maccone, and Peter W. Shor. Entanglement assisted capacity of the broadband lossy channel. *Physical Review Letters*, 91(4):047901, July 2003.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[108] Roy J. Glauber. Coherent and incoherent states of the radiation field. *Physical Review*, 131(6):2766–2788, Sep 1963.

[109] Roy J. Glauber. The quantum theory of optical coherence. *Physical Review*, 130(6):2529–2539, Jun 1963.

[110] Roy J. Glauber. One hundred years of light quanta. In Karl Grandin, editor, *Les Prix Nobel. The Nobel Prizes 2005*, pages 90–91. Nobel Foundation, 2005.

[111] J. P. Gordon. Noise at optical frequencies; information theory. In P. A. Miles, editor, *Quantum Electronics and Coherent Light; Proceedings of the International School of Physics Enrico Fermi, Course XXXI*, pages 156–181, Academic Press New York, 1964.

[112] Daniel Gottesman. Class of quantum error-correcting codes saturating the quantum Hamming bound. *Physical Review A*, 54(3):1862–1868, September 1996.

[113] Daniel Gottesman. *Stabilizer Codes and Quantum Error Correction*. PhD thesis, California Institute of Technology (arXiv:quant-ph/9705052), 1997.

[114] Markus Grassl, Thomas Beth, and Thomas Pellizzari. Codes for the quantum erasure channel. *Physical Review A*, 56(1):33–38, July 1997.

[115] Brian Greene. *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*. W. W. Norton & Company, 1999.

[116] David J. Griffiths. *Introduction to Quantum Mechanics*. Prentice-Hall, Inc., 1995.

[117] Berry Groisman, Sandu Popescu, and Andreas Winter. Quantum, classical, and total amount of correlations in a quantum state. *Physical Review A*, 72(3):032317, September 2005.

[118] Saikat Guha and Jeffrey H. Shapiro. Classical information capacity of the bosonic broadcast channel. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 1896–1900, Nice, France, June 2007.

[119] Saikat Guha, Jeffrey H. Shapiro, and Baris I. Erkmen. Classical capacity of bosonic broadcast communication and a minimum output entropy conjecture. *Physical Review A*, 76(3):032303, September 2007.

[120] Saikat Guha, Jeffrey H. Shapiro, and Baris I. Erkmen. Capacity of the bosonic wiretap channel and the entropy photon-number inequality. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 91–95, Toronto, Ontario, Canada, July 2008. arXiv:0801.0841.

[121] Mitsuru Hamada. Information rates achievable with algebraic codes on quantum discrete memoryless channels. *IEEE Transactions in Information Theory*, 51(12):4263–4277, December 2005. arXiv:quant-ph/0207113.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[122] Jim Harrington and John Preskill. Achievable rates for the Gaussian quantum channel. *Physical Review A*, 64(6):062301, November 2001.

[123] Aram Harrow. Coherent communication of classical messages. *Physical Review Letters*, 92:097902, March 2004.

[124] Aram W. Harrow and Hoi-Kwong Lo. A tight lower bound on the classical communication cost of entanglement dilution. *IEEE Transactions on Information Theory*, 50(2):319–327, February 2004.

[125] Matthew B. Hastings. Superadditivity of communication capacity using entangled inputs. *Nature Physics*, 5:255–257, April 2009. arXiv:0809.3972.

[126] Paul Hausladen, Richard Jozsa, Benjamin Schumacher, Michael Westmoreland, and William K. Wootters. Classical information capacity of a quantum channel. *Physical Review A*, 54(3):1869–1876, September 1996.

[127] Paul Hausladen, Benjamin Schumacher, Michael Westmoreland, and William K. Wootters. Sending classical bits via quantum its. *Annals of the New York Academy of Sciences*, 755:698–705, April 1995.

[128] Masahito Hayashi. *Quantum Information: An Introduction*. Springer, 2006.

[129] Masahito Hayashi and Keiji Matsumoto. Variable length universal entanglement concentration by local operations and its application to teleportation and dense coding. September 2001. arXiv:quant-ph/0109028.

[130] Masahito Hayashi and Hiroshi Nagaoka. General formulas for capacity of classical-quantum channels. *IEEE Transactions on Information Theory*, 49(7):1753–1768, 2003.

[131] Patrick Hayden. The maximal p-norm multiplicativity conjecture is false. July 2007. arXiv:0707.3291.

[132] Patrick Hayden, M. Horodecki, Andreas Winter, and Jon Yard. A decoupling approach to the quantum capacity. *Open Systems & Information Dynamics*, 15:7–19, March 2008.

[133] Patrick Hayden, Richard Jozsa, and Andreas Winter. Trading quantum for classical resources in quantum data compression. *Journal of Mathematical Physics*, 43(9):4404–4444, 2002.

[134] Patrick Hayden, Peter W. Shor, and Andreas Winter. Random quantum codes from Gaussian ensembles and an uncertainty relation. *Open Systems & Information Dynamics*, 15:71–89, March 2008.

[135] Patrick Hayden and Andreas Winter. Communication cost of entanglement transformations. *Physical Review A*, 67(1):012326, January 2003.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[136] Patrick Hayden and Andreas Winter. Counterexamples to the maximal p-norm multiplicativity conjecture for all p > 1. *Communications in Mathematical Physics*, 284(1):263–280, 2008. arXiv:0807.4753.

[137] Werner Heisenberg. Über quantentheoretische umdeutung kinematischer und mechanischer beziehungen. *Zeitschrift für Physik*, 33:879–893, 1925.

[138] Carl W. Helstrom. Quantum detection and estimation theory. *Journal of Statistical Physics*, 1:231–252, 1969.

[139] Carl W. Helstrom. *Quantum Detection and Estimation Theory*. Academic, New York, 1976.

[140] Nick Herbert. Flash—a superluminal communicator based upon a new kind of quantum measurement. *Foundations of Physics*, 12(12):1171–1179, January 1982.

[141] A. S. Holevo and R. F. Werner. Evaluating capacities of bosonic Gaussian channels. *Physical Review A*, 63(3):032312, February 2001.

[142] Alexander S. Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problems of Information Transmission*, 9:177–183, 1973.

[143] Alexander S. Holevo. Statistical problems in quantum physics. In *Second Japan-USSR Symposium on Probability Theory*, volume 330 of *Lecture Notes in Mathematics*, pages 104–119. Springer Berlin / Heidelberg, 1973.

[144] Alexander S. Holevo. The capacity of the quantum channel with general signal states. *IEEE Transactions on Information Theory*, 44:269–273, 1998.

[145] Alexander S. Holevo. *An Introduction to Quantum Information Theory*. Moscow Center of Continuous Mathematical Education, Moscow, 2002. In Russian.

[146] Alexander S. Holevo. On entanglement assisted classical capacity. *Journal of Mathematical Physics*, 43(9):4326–4333, 2002.

[147] M. Horodecki. Limits for compression of quantum information carried by ensembles of mixed states. *Physical Review A*, 57(5):3364–3369, May 1998.

[148] M. Horodecki, Jonathan Oppenheim, and Andreas Winter. Partial quantum information. *Nature*, 436:673–676, 2005.

[149] M. Horodecki, Jonathan Oppenheim, and Andreas Winter. Quantum state merging and negative information. *Communications in Mathematical Physics*, 269:107–136, 2007.

[150] M. Horodecki, Peter W. Shor, and Mary Beth Ruskai. Entanglement breaking channels. *Reviews in Mathematical Physics*, 15(6):629–641, 2003. arXiv:quant-ph/0302031.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[151] Michal Horodecki, Pawel Horodecki, and Ryszard Horodecki. Separability of mixed states: necessary and sufficient conditions. *Physics Letters A*, 223(1-2):1–8, November 1996.

[152] Michal Horodecki, Pawel Horodecki, Ryszard Horodecki, Debbie Leung, and Barbara Terhal. Classical capacity of a noiseless quantum channel assisted by noisy entanglement. *Quantum Information and Computation*, 1(3):70–78, 2001. arXiv:quant-ph/0106080.

[153] Pawel Horodecki. Separability criterion and inseparable mixed states with positive partial transposition. *Physics Letters A*, 232(5):333–339, 1997.

[154] R. Horodecki and P. Horodecki. Quantum redundancies and local realism. *Physics Letters A*, 194:147–152, 1994.

[155] Ryszard Horodecki, P. Horodecki, M. Horodecki, and Karol Horodecki. Quantum entanglement. *Reviews of Modern Physics*, 81(2):865–942, June 2009.

[156] Min-Hsiu Hsieh, Igor Devetak, and Andreas Winter. Entanglement-assisted capacity of quantum multiple-access channels. *IEEE Transactions on Information Theory*, 54(7):3078–3090, 2008.

[157] Min-Hsiu Hsieh, Zhicheng Luo, and Todd Brun. Secret-key-assisted private classical communication capacity over quantum channels. *Physical Review A*, 78(4):042306, October 2008.

[158] Min-Hsiu Hsieh and Mark M. Wilde. Public and private communication with a quantum channel and a secret key. *Physical Review A*, 80(2):022306, August 2009.

[159] Min-Hsiu Hsieh and Mark M. Wilde. Entanglement-assisted communication of classical and quantum information. *IEEE Transactions on Information Theory*, 56(9):4682–4704, September 2010.

[160] Min-Hsiu Hsieh and Mark M. Wilde. Trading classical communication, quantum communication, and entanglement in quantum Shannon theory. *IEEE Transactions on Information Theory*, 56(9):4705–4730, September 2010.

[161] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620, 1957.

[162] E. T. Jaynes. Information theory and statistical mechanics II. *Physical Review*, 108:171, 1957.

[163] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[164] Tomas Jochym-O'Connor, Kamil Brádler, and Mark M. Wilde. Trade-off coding for universal qudit cloners motivated by the Unruh effect. *Journal of Physics A*, 44:415306, March 2011. arXiv:1103.0286.

[165] Richard Jozsa. Fidelity for mixed quantum states. *Journal of Modern Optics*, 41(12):2315–2323, 1994.

[166] Richard Jozsa, M. Horodecki, Paweł Horodecki, and Ryszard Horodecki. Universal quantum information compression. *Physical Review Letters*, 81(8):1714–1717, August 1998.

[167] Richard Jozsa and Stuart Presnell. Universal quantum information compression and degrees of prior knowledge. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 459(2040):3061–3077, December 2003.

[168] Richard Jozsa and Benjamin Schumacher. A new proof of the quantum noiseless coding theorem. *Journal of Modern Optics*, 41(12):2343–2349, December 1994.

[169] Phillip Kaye and Michele Mosca. Quantum networks for concentrating entanglement. *Journal of Physics A: Mathematical and General*, 34(35):6939, August 2001.

[170] Christopher King. Additivity for unital qubit channels. *Journal of Mathematical Physics*, 43(10):4641–4653, 2002. arXiv:quant-ph/0103156.

[171] Christopher King. The capacity of the quantum depolarizing channel. *IEEE Transactions on Information Theory*, 49(1):221–229, January 2003.

[172] Christopher King, Keiji Matsumoto, Michael Nathanson, and Mary Beth Ruskai. Properties of conjugate channels with applications to additivity and multiplicativity. *Markov Processes and Related Fields*, 13(2):391–423, 2007. J. T. Lewis memorial issue.

[173] Alexei Y. Kitaev. *Uspekhi Mat. Nauk.*, 52(53), 1997.

[174] Rochus Klesse. A random coding based proof for the quantum coding theorem. *Open Systems & Information Dynamics*, 15:21–45, March 2008.

[175] Emanuel H. Knill, Raymond Laflamme, and Wojciech H. Zurek. Resilient quantum computation. *Science*, 279:342–345, 1998. quant-ph/9610011.

[176] Masato Koashi and Nobuyuki Imoto. Teleportation cost and hybrid compression of quantum signals. *arXiv:quant-ph/0104001*, 2001.

[177] Robert Koenig, Renato Renner, and Christian Schaffner. The operational meaning of min- and max-entropy. *IEEE Transactions on Information Theory*, 55(9):4337–4347, September 2009. arXiv:0807.1338.

[178] Isaac Kremsky, Min-Hsiu Hsieh, and Todd A. Brun. Classical enhancement of quantum-error-correcting codes. *Physical Review A*, 78(1):012341, 2008.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[179] Greg Kuperberg. The capacity of hybrid quantum memory. *IEEE Transactions on Information Theory*, 49(6):1465–1473, June 2003.

[180] Raymond Laflamme, Cesar Miquel, Juan Pablo Paz, and Wojciech Hubert Zurek. Perfect quantum error correcting code. *Physical Review Letters*, 77(1):198–201, July 1996.

[181] Rolf Landauer. Is quantum mechanics useful? *Philosophical Transactions of the Royal Society: Physical and Engineering Sciences*, 353(1703):367–376, December 1995.

[182] L. B. Levitin. On the quantum measure of information. In *Proceedings of the Fourth All-Union Conference on Information and Coding Theory, Sec. II,*, Tashkent, 1969.

[183] Ke Li, Andreas Winter, XuBo Zou, and Guang-Can Guo. Private capacity of quantum channels is not additive. *Physical Review Letters*, 103(12):120501, Sep 2009.

[184] Elliott H. Lieb and Mary Beth Ruskai. Proof of the strong subadditivity of quantum-mechanical entropy. *Journal of Mathematical Physics*, 14:1938–1941, 1973.

[185] Seth Lloyd. Capacity of the noisy quantum channel. *Physical Review A*, 55(3):1613–1622, March 1997.

[186] Hoi-Kwong Lo. Quantum coding theorem for mixed states. *Optics Communications*, 119(5-6):552–556, September 1995.

[187] Hoi-Kwong Lo and Sandu Popescu. Classical communication cost of entanglement manipulation: Is entanglement an interconvertible resource? *Physical Review Letters*, 83(7):1459–1462, August 1999.

[188] Hoi-Kwong Lo and Sandu Popescu. Concentrating entanglement by local actions: Beyond mean values. *Physical Review A*, 63(2):022301, January 2001.

[189] David MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, September 2003.

[190] J. P. McEvoy and Oscar Zarate. *Introducing Quantum Theory*. Totem Books, third edition, October 2004.

[191] Charles W. Misner, Kip S. Thorne, and Wojciech H. Zurek. John Wheeler, relativity, and quantum information. *Physics Today*, April 2009.

[192] Milán Mosonyi and Nilanjana Datta. Generalized relative entropies and the capacity of classical-quantum channels. *Journal of Mathematical Physics*, 50(7):072104, 2009.

[193] Justin Mullins. The topsy turvy world of quantum computing. *IEEE Spectrum*, 38(2):42–49, February 2001.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[194] Michael A. Nielsen. *Quantum information theory*. PhD thesis, University of New Mexico, 1998. arXiv:quant-ph/0011036.

[195] Michael A. Nielsen. Conditions for a class of entanglement transformations. *Physical Review Letters*, 83(2):436–439, July 1999.

[196] Michael A. Nielsen. A simple formula for the average gate fidelity of a quantum dynamical operation. *Physics Letters A*, 303(4):249 – 252, 2002.

[197] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[198] Michael A. Nielsen and Denés Petz. A simple proof of the strong subadditivity inequality. August 2004. arXiv:quant-ph/0408130.

[199] Tomohiro Ogawa and Hiroshi Nagaoka. Making good codes for classical-quantum channel coding via quantum hypothesis testing. *IEEE Transactions on Information Theory*, 53(6):2261–2266, June 2007.

[200] Masanori Ohya and Denes Petz. *Quantum Entropy and Its Use*. Springer, 1993.

[201] Masanao Ozawa. Quantum measuring processes of continuous observables. *Journal of Mathematical Physics*, 25(1):79–87, 1984.

[202] Arun Kumar Pati and Samuel L. Braunstein. Impossibility of deleting an unknown quantum state. *Nature*, 404:164–165, March 2000.

[203] Asher Peres. How the no-cloning theorem got its name. 2002. arXiv:quant-ph/0205076.

[204] John R. Pierce. The early days of information theory. *IEEE Transactions on Information Theory*, IT-19(1):3–8, January 1973.

[205] Max Planck. Ueber das gesetz der energieverteilung im normalspectrum. *Annalen der Physik*, 4:553–563, 1901.

[206] John Preskill. Reliable quantum computers. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 454(1969):385–410, January 1998.

[207] Renato Renner. *Security of Quantum Key Distribution*. PhD thesis, ETH Zurich, September 2005. arXiv:quant-ph/0512258.

[208] R. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2):120–126, 1978.

[209] J. J. Sakurai. *Modern Quantum Mechanics (2nd Edition)*. Addison Wesley, January 1994.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[210] Ivan Savov. Distributed compression and squashed entanglement. Master's thesis, McGill University, February 2008. arXiv:0802.0694.

[211] Ivan Savov. *Network information theory for classical-quantum channels*. PhD thesis, McGill University, 2012.

[212] Valerio Scarani, Helle Bechmann-Pasquinucci, Nicolas J. Cerf, Miloslav Dušek, Norbert Lütkenhaus, and Momtchil Peev. The security of practical quantum key distribution. *Reviews of Modern Physics*, 81(3):1301–1350, September 2009.

[213] Valerio Scarani, Sofyan Iblisdir, Nicolas Gisin, and Antonio Acín. Quantum cloning. *Reviews of Modern Physics*, 77(4):1225–1256, November 2005.

[214] Erwin Schrödinger. Quantisierung als eigenwertproblem. *Annalen der Physik*, 79:361–376, 1926.

[215] Erwin Schrödinger. Discussion of probability relations between separated systems. *Proceedings of the Cambridge Philosophical Society*, 31:555–563, 1935.

[216] Benjamin Schumacher. Quantum coding. *Physical Review A*, 51(4):2738–2747, April 1995.

[217] Benjamin Schumacher. Sending entanglement through noisy quantum channels. *Physical Review A*, 54(4):2614–2628, October 1996.

[218] Benjamin Schumacher and Michael A. Nielsen. Quantum data processing and error correction. *Physical Review A*, 54(4):2629–2635, October 1996.

[219] Benjamin Schumacher and Michael D. Westmoreland. Sending classical information via noisy quantum channels. *Physical Review A*, 56(1):131–138, July 1997.

[220] Benjamin Schumacher and Michael D. Westmoreland. Quantum privacy and quantum coherence. *Physical Review Letters*, 80(25):5695–5697, June 1998.

[221] Benjamin Schumacher and Michael D. Westmoreland. Approximate quantum error correction. *Quantum Information Processing*, 1(1/2):5–12, 2002.

[222] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[223] Peter W. Shor. Algorithms for quantum computation: Discrete logarithms and factoring. In *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, pages 124–134, Los Alamitos, California, 1994. IEEE Computer Society Press.

[224] Peter W. Shor. Scheme for reducing decoherence in quantum computer memory. *Physical Review A*, 52(4):R2493–R2496, October 1995.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[225] Peter W. Shor. Fault-tolerant quantum computation. *Annual IEEE Symposium on Foundations of Computer Science*, page 56, 1996.

[226] Peter W. Shor. Additivity of the classical capacity of entanglement-breaking quantum channels. *Journal of Mathematical Physics*, 43(9):4334–4340, 2002. arXiv:quant-ph/0201149.

[227] Peter W. Shor. The quantum channel capacity and coherent information. In *Lecture Notes, MSRI Workshop on Quantum Computation*, 2002.

[228] Peter W. Shor. Equivalence of additivity questions in quantum information theory. *Communications in Mathematical Physics*, 246(3):453–472, 2004. arXiv:quant-ph/0305035.

[229] Peter W. Shor. *Quantum Information, Statistics, Probability (Dedicated to A. S. Holevo on the occasion of his 60th Birthday): The classical capacity achievable by a quantum channel assisted by limited entanglement*. Rinton Press, Inc., 2004. arXiv:quant-ph/0402129.

[230] Graeme Smith. Private classical capacity with a symmetric side channel and its application to quantum cryptography. *Physical Review A*, 78(2):022306, August 2008.

[231] Graeme Smith, Joseph M. Renes, and John A. Smolin. Structured codes improve the Bennett-Brassard-84 quantum key rate. *Physical Review Letters*, 100(17):170502, 2008.

[232] Graeme Smith and John A. Smolin. Degenerate quantum codes for Pauli channels. *Physical Review Letters*, 98(3):030501, January 2007.

[233] Graeme Smith, John A. Smolin, and Jon Yard. Quantum communication with gaussian channels of zero quantum capacity. *Nature Photonics*, 5:624–627, August 2011. arXiv:1102.4580.

[234] Graeme Smith and Jon Yard. Quantum communication with zero-capacity channels. *Science*, 321:1812–1815, September 2008.

[235] Andrew M. Steane. Error correcting codes in quantum theory. *Physical Review Letters*, 77(5):793–797, July 1996.

[236] W. F. Stinespring. Positive functions on C*-algebras. *Proceedings of the American Mathematical Society*, 6:211–216, 1955.

[237] Marco Tomamichel, Roger Colbeck, and Renato Renner. A fully quantum asymptotic equipartition property. *IEEE Transactions on Information Theory*, 55(12):5840—5847, December 2009. arXiv:0811.1221.

[238] Marco Tomamichel, Roger Colbeck, and Renato Renner. Duality between smooth min- and max-entropies. *IEEE Transactions on Information Theory*, 56(9):4674–4681, September 2010. arXiv:0907.5238.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[239] Marco Tomamichel and Renato Renner. Uncertainty relation for smooth entropies. *Physical Review Letters*, 106:110506, March 2011. arXiv:1009.2015.

[240] Armin Uhlmann. The "transition probability" in the state space of a *-algebra. *Reports on Mathematical Physics*, 9(2):273–279, 1976.

[241] William G. Unruh. Maintaining coherence in quantum computers. *Physical Review A*, 51(2):992–997, Feb 1995.

[242] Dennis von Kretschmann. *Information Transfer through Quantum Channels*. PhD thesis, Technische Universität Braunschweig, 2007.

[243] John von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, October 1996.

[244] Ligong Wang and Renato Renner. One-shot classical-quantum capacity and hypothesis testing. *Accepted into Physical Review Letters*, July 2010. arXiv:1007.5456.

[245] Stephanie Wehner and Andreas Winter. Entropic uncertainty relations—a survey. *New Journal of Physics*, 12:025009, February 2010. arXiv:0907.3704.

[246] A. Wehrl. General properties of entropy. *Reviews of Modern Physics*, 50:221–260, 1978.

[247] Stephen Wiesner. Conjugate coding. *SIGACT News*, 15(1):78–88, 1983.

[248] Mark M. Wilde. Comment on "Secret-key-assisted private classical communication capacity over quantum channels". *Physical Review A*, 83(4):046303, April 2011.

[249] Mark M. Wilde and Todd A. Brun. Unified quantum convolutional coding. In *Proceedings of the IEEE International Symposium on Information Theory*, Toronto, Ontario, Canada, July 2008. arXiv:0801.0821.

[250] Mark M. Wilde, Patrick Hayden, and Saikat Guha. Information trade-offs for optical quantum communication. *Physical Review Letters*, 108(14):140501, April 2012. arXiv:1105.0119.

[251] Mark M. Wilde and Min-Hsiu Hsieh. Entanglement generation with a quantum channel and a shared state. *Proceedings of the 2010 IEEE International Symposium on Information Theory*, pages 2713–2717, June 2010. arXiv:0904.1175.

[252] Mark M. Wilde and Min-Hsiu Hsieh. Public and private resource trade-offs for a quantum channel. *Accepted into Quantum Information Processing*, 2010. arXiv:1005.3818.

[253] Mark M. Wilde and Min-Hsiu Hsieh. The quantum dynamic capacity formula of a quantum channel. *Accepted into Quantum Information Processing*, April 2010. arXiv:1004.0458.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[254] Mark M. Wilde, Hari Krovi, and Todd A. Brun. Coherent communication with continuous quantum variables. *Physical Review A*, 75(6):060303(R), 2007.

[255] Andreas Winter. Coding theorem and strong converse for quantum channels. *IEEE Transactions on Information Theory*, 45(7):2481–2485, 1999.

[256] Andreas Winter. *Coding Theorems of Quantum Information Theory*. PhD thesis, Universität Bielefeld, July 1999. arXiv:quant-ph/9907077.

[257] Andreas Winter. The capacity of the quantum multiple access channel. *IEEE Transactions on Information Theory*, 47:3059–3065, 2001.

[258] Andreas Winter. The maximum output p-norm of quantum channels is not multiplicative for any p $>$ 2. July 2007. arXiv:0707.0402.

[259] Andreas Winter and Serge Massar. Compression of quantum-measurement operations. *Physical Review A*, 64(1):012311, June 2001.

[260] Andreas J. Winter. "Extrinsic" and "intrinsic" data in quantum measurements: asymptotic convex decomposition of positive operator valued measures. *Communications in Mathematical Physics*, 244(1):157–185, 2004.

[261] Michael M. Wolf, Toby S. Cubitt, and David Perez-Garcia. Are problems in quantum information theory (un)decidable? November 2011. arXiv:1111.5425.

[262] Michael M. Wolf and David Pérez-García. Quantum capacities of channels with small environment. *Physical Review A*, 75(1):012303, January 2007.

[263] Michael M. Wolf, David Pérez-García, and Geza Giedke. Quantum capacities of bosonic channels. *Physical Review Letters*, 98(13):130501, March 2007.

[264] Jacob Wolfowitz. *Coding theorems of information theory*. Springer-Verlag, 1978.

[265] William K. Wootters and Wojciech H. Zurek. A single quantum cannot be cloned. *Nature*, 299:802–803, 1982.

[266] Jon Yard. *Simultaneous classical-quantum capacities of quantum multiple access channels*. PhD thesis, Stanford University, Stanford, CA, 2005. arXiv:quant-ph/0506050.

[267] Jon Yard and Igor Devetak. Optimal quantum source coding with quantum side information at the encoder and decoder. *IEEE Transactions in Information Theory*, 55(11):5339–5351, November 2009. arXiv:0706.2907.

[268] Jon Yard, Igor Devetak, and Patrick Hayden. Capacity theorems for quantum multiple access channels. In *Proceedings of the International Symposium on Information Theory*, pages 884–888, Adelaide, Australia, September 2005.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

[269] Jon Yard, Patrick Hayden, and Igor Devetak. Capacity theorems for quantum multiple-access channels: Classical-quantum and quantum-quantum capacity regions. *IEEE Transactions on Information Theory*, 54(7):3091–3113, 2008.

[270] Jon Yard, Patrick Hayden, and Igor Devetak. Quantum broadcast channels. *IEEE Transactions on Information Theory*, 57(10):7147–7162, October 2011. arXiv:quant-ph/0603098.

[271] Ming-Yong Ye, Yan-Kui Bai, and Z. D. Wang. Quantum state redistribution based on a generalized decoupling. *Physical Review A*, 78(3):030302, September 2008.

[272] Brent J. Yen and Jeffrey H. Shapiro. Multiple-access bosonic communications. *Physical Review A*, 72(6):062312, December 2005.

[273] Raymond W. Yeung. *A First Course in Information Theory*. Information Technology: Transmission, Processing, and Storage. Springer (Kluwer Academic/Plenum Publishers), New York, New York, USA, March 2002.

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Index

accessible information, 265, 284
    of a channel, 470
Alicki-Fannes' inequality, 298
amplitude damping channel, 142, 519, 582
anticommutator, 78
asymptotic equipartition theorem, 47, 343

Bell inequality, 24, 98
Bell measurement, 178
Bell states, 100, 106, 175, 178
binary entropy, 25, 249
binary symmetric channel, 49
Bloch sphere, 70, 115
Born rule, 69, 82, 95, 111
bosonic channel, 621

classical capacity theorem, 474
classical-quantum state, 131, 149
coherent bit channel, 194
coherent communication, 525
coherent communication identity, 202, 530, 631
coherent information, 267, 280, 567
    conditional, 281
    convexity, 286
    of a channel, 329
      additivity for a degradable channel, 330
      positivity, 330
      superadditivity, 584
commutator, 78
complementary channel, 158, 163
completely positive trace-preserving map, 135
conditional quantum channel, 149
conditional typicality, 58
covering lemma, 417

data processing inequality, 259
decoupling approach, 593
density operator, 110, 112
dephasing channel, 139, 521, 618
    generalized, 163
depolarizing channel, 140, 483, 521, 584

ebit, 175
elementary coding, 177
entanglement concentration, 445, 629
entanglement dilution, 629
entanglement distillation, 589
entanglement distribution, 171, 174, 187
entanglement swapping, 183
entanglement-assisted
    classical communication, 30, 185, 493, 630
      with feedback, 512
    coherent communication, 527
    quantum communication, 529, 602, 631
entanglement-breaking channel, 144
erasure channel, 143, 517, 570, 581
expurgation, 57, 414, 560
extension, 155

Fannes' Inequality, 300
Fannes-Audenaert inequality, 300
Fano's inequality, 261
fidelity, 28, 226
    entanglement fidelity, 241
    expected, 227
    joint concavity, 231
    monotonicity, 232
    Uhlmann, 228

Fourier transform, 103

gentle measurement, 237
coherent, 240
for ensembles, 239

Hadamard channel, 164, 480, 596, 616
Hadamard gate, 79
Heisenberg-Weyl operators, 102, 501
Hilbert-Schmidt distance, 244
Holevo bound, 32, 265, 297, 404, 493
Holevo information, 284
of a channel, 317
concavity, 322, 323
superadditivity, 487
HSW theorem, 29, 403, 474, 600, 630

indeterminism, 22
indicator function, 120
information content, 41, 248
isometric extension, 156

Kraus operators, 134
Kraus representation, 110

law of large numbers, 45, 343
LSD theorem, 30, 630

Markov's inequality, 54, 414
matrix representation, 76
mutual information, 60, 254
conditional, 256
of a classical channel, 308
additivity, 311

network quantum Shannon theory, 633
no-cloning theorem, 33, 93, 549, 567, 570
no-deletion theorem, 94
noisy channel coding theorem, 52

observable, 82
Operator Chernoff Bound, 422, 638

packing lemma, 403, 476, 503, 533
partial trace, 128
Pauli channel, 140
Pauli matrices, 78
polar decomposition, 636
positive operator-valued measure, 110, 123
private classical capacity
superadditivity, 564
private classical capacity theorem, 552
private classical communication, 549, 630
secret-key assisted, 564
private information
of a quantum channel, 334
additivity for a degradable channel, 337
of a wiretap channel, 314
additivity for degraded channels, 315
positivity, 315
suparadditivity, 563
probability amplitudes, 69
purification, 154

quantum capacity theorem, 18, 30, 185, 567, 601
quantum data processing inequality, 294
quantum dynamic capacity formula, 611
quantum dynamic capacity theorem, 595, 598
quantum hypothesis testing, 222
quantum information source, 380, 433
quantum instrument, 147
coherent, 165
quantum interference, 22, 23
quantum key distribution, 33
quantum mutual information, 282
chain rule, 285
conditional
positivity, 285
of a channel, 324
additivity, 324
positivity, 282
quantum relative entropy, 286
conditions for infinite value, 288
joint convexity, 291
monotonicity, 289, 643
positivity, 287
quantum reverse Shannon theorem, 632

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

quantum teleportation, 24, 29, 34, 171, 179, 189, 206
coherent, 200
noisy, 186, 631
quantum typicality, 379
conditional, 389
strong, 393
weak, 392
qudit, 102, 186

relative entropy, 255
positivity, 258
remote state preparation, 182, 633

sample entropy, 45, 345, 346
conditional, 354
joint, 351
Schmidt decomposition, 107, 501
Schumacher compression, 28, 269, 433, 629
separable states, 110, 125
Shannon compression, 42, 349
Shannon entropy, 42, 248
concavity, 251
conditional, 252
joint, 253
positivity, 250
square-root measurement, 409, 477
state merging, 36, 280
state redistribution, 632
state transfer
classical-assisted, 537, 631
coherent, 535
quantum-assisted, 536, 631
Stinespring dilation, 158
strong subadditivity, 285, 286
super-dense coding, 171, 178, 188
coherent, 198
noisy, 532, 631
superactivation, 36, 568, 587
superposition, 22, 23, 69, 83

tensor product, 88
time-sharing, 539
trace distance, 28, 218, 219
monotonicity, 225
operational interpretation, 222
triangle inequality, 224
trace norm, 218
trade-off coding, 538, 631
twirling, 141
type, 358
bound on number of types, 359
type class, 359
projector, 400
subspace, 400, 451, 501
typical, 363
typical projector, 382
conditionally
strong, 394
typical sequence, 46, 343, 346
jointly, 351
typical set, 346, 347
conditionally, 354
jointly, 352
strongly, 358
typical subspace, 28, 382
conditionally
strong, 394
weak, 392
measurement, 383, 437, 451
typicality
strong, 356
conditional, 367
joint, 366
weak, 345
conditional, 354
joint, 351

uncertainty principle, 86
unit resource capacity region, 205, 595

von Neumann entropy, 267, 268
additivity, 274
concavity, 270, 284
conditional, 278
concavity, 286
continuity, 298

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

joint, 272
operational interpretation, 434, 445
positivity, 270
subadditivity, 288
von Neumann measurement, 84

©2012 Mark M. Wilde—This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License